**Human Aspects of Nuclear Safety**
**NEA/CSNI/R(2016)17/ADD1**
March 2017
www.oecd-nea.org

# Human Factors Validation of Nuclear Power Plant Control Room Designs and Modifications

Proceedings of the Expert Workshop
Charlotte, United States
19-21 February 2015

## Appendix B

OECD
BETTER POLICIES FOR BETTER LIVES

NEA
NUCLEAR ENERGY AGENCY

# APPENDIX B – WHITE PAPERS

The white papers in this appendix were solicited by the workshop organizing committee and submitted by the workshop participants in advance of the workshop as a means to facilitate the development and sharing of participant views on topics related to the workshop theme of achieving reasonable confidence in the validation of main control room systems. The white papers were made available via a workshop website for participants to read and consider as they prepared for their participation in the workshop.

## Appendix B - Table of Contents

For the readers convenience the papers in this appendix have been organized according to their relationship to the Challenge Questions addressed in this workshop. It should be noted that these papers were not submitted in response to these Challenge Questions but were judged to be relevant to them.

***What are the critical considerations in defining the objectives of a control room validation and how do these impacts achieving reasonable confidence?***

***What methods, approaches, resources, or rationales might be used for deriving performance requirements, selecting measures, and establishing acceptance criteria so as to support reasonable confidence?***

***What methods might be used to develop scenarios that maximize the amount and relevance of information in support of the validation conclusions and achieving reasonable confidence?***

***How should the validation results be aggregated and analyzed to determine the final validation conclusions? Are inferential statistics meaningful in the context of MCR validation? If not, why not? What substitutes might be proposed as alternatives to traditional statistical modeling approaches?***

***General***..........

## References
## (The following papers, published elsewhere, were provided for consideration by the workshop participants

Braarud, P. Ø. (2015). An Overall Framework for the Definition of Requirements, Criteria and Human Engineering Discrepancies for Control Room Validation. In Proceedings of the Ninth American Nuclear Society International Topical on Nuclear Plant Instrumentation, Controls and Human-Machine Interface Technologies (NPIC & HMIT 2015). La Grange, IL: American Nuclear Society.

Braaurd, P., Eithreim, M. & Fernandes, A. (2015). “SCORE” – An Integrated Performance Measure for Control Room Validation. In Proceedings of the Ninth American Nuclear Society International Topical on Nuclear Plant Instrumentation, Controls and Human-Machine Interface Technologies (NPIC & HMIT 2015). La Grange, IL: American Nuclear Society.

Elsner, T., Freitag, T. & Rivere, C. (2015). Lessons Learned during the Human Factors Guided Validation of Human-System Interfaces in New Build NPPs and Modernization Projects. In Proceedings of the Ninth American Nuclear Society International Topical on Nuclear Plant Instrumentation, Controls and Human-Machine Interface Technologies (NPIC & HMIT 2015). La Grange, IL: American Nuclear Society.

RIVERE, C. (2015). Human Factors Engineering Verification and Validation Process of New Nuclear Power Plant Control Room: How to Bridge the Gap from Stepwise V&V to final ISV. In Proceedings of the Ninth American Nuclear Society International Topical on Nuclear Plant Instrumentation, Controls and Human-Machine Interface Technologies (NPIC & HMIT 2015). La Grange, IL: American Nuclear Society.

Skraaning, G. & Strand, S. (2015). Integrated System Validation: The Acceptability Analysis Process. In Proceedings of the Ninth American Nuclear Society International Topical on Nuclear Plant Instrumentation, Controls and Human-Machine Interface Technologies (NPIC & HMIT 2015). La Grange, IL: American Nuclear Society.

# TECNATOM PERSPECTIVE IN RELATION TO VALIDATION AND ISV FROM THE HFE VIEWPOINT FOR NUCLEAR POWER PLANTS

Pedro Trueba
Alonso Tecnatom,
S.A.
Avda. Montes de Oca 1, San Sebastián de los Reyes, 28703 Madrid, Spain

## 1 Introduction

The word validation has many different meanings depending on the context where it applies. In a general way validation is a testing intended to ensure a product, service, or system (or portion of a system) meets the operational needs of the final user.

Validation and Integrated System Validation (ISV) do not have necessarily the same meaning, because the latter has higher requirements in the sense that what must be tested must be integrated with other products, services, systems or parts.

In the nuclear field ISV is an activity associated to the area of Human Factors Engineering (HFE), being mainly related to the Main Control Room (MCR) and Remote Shutdown Stations (RSS). According to the HFE Model included in NUREG-0711, ISV is defined as an evaluation, using performance-based tests, to determine whether an integrated system's design (i.e., hardware, software, and personnel elements) meets performance requirements and supports the plant's safe operation.

The considerations described in this paper will focus in the different types of validations that can exist, related to the nuclear field, and with special emphasis in the most challenging validation, the ISV, as the final stage of a HFE process (for new designs and for certain types of design modifications), based on the experience of Tecnatom.

## 2 Validation Cases

Following international standards such as NUREG-0711, IEC-60964/61771 or ISO-11064, a validation and/or an ISV can be considered applicable for new designs and design modifications, and where three different cases can be considered:

Case 1: Validation/ISV for Research and Development (R&D) activities, with the objective of testing new technological approaches and human performance
Case 2: Validation/ISV for Licensing activities associated to a generic new or modified design
Case 3 Validation/ISV for Licensing activities associated to a specific site for a new or modified design

The above listed cases can have a common objective related to confirm plant safety with any design under consideration, but can have significant differences in validation objectives, methodological approach, personnel involved, duration and execution time constraints.

There are a wide variety of validation methods and approaches that can be found in international literature for applying them as part of any validation process. All approaches can be compatible with the execution time and schedule expected for Case 1, but can have certain applicability limitations for Cases 2 and 3 listed above. Consequently the context is the very first issue to take into account when planning a validation or an ISV.

### 2.1 Case 1: Validations as Part of R&D Activities

For Case 1, almost all possible validation methods can be applicable, because it is more associated to early stages of a new design development, testing new technologies or testing of human performance for HFE Studies or Experiments.

For Case 1 related to fully new designs, the possibilities of having an integrated design for validation are low as not all the Human System Interface (HSI) may be designed or implemented. ISV in this case is only possible taking advantage of previous/similar designs and facilities.

Another approach for Case 1 is the execution of partial validations, where partial could mean that there is no complete availability of all HSI components, no complete training, no complete simulator functionality, or tests only focused on specific plant conditions (relevant to what wants to be validated).

In the case of Human Performance studies, ISV is more feasible because it normally needs the introduction of new HSI components in an existing operational concept or control room. However laboratories are normally used for this purpose that do not resemble necessarily a real MCR HSI. The approach for these studies are to analyze particular human abilities like situation awareness or workload, and where the plant conditions being simulated represented need not be realistic (like freezing the test to gather information or introducing non-realistic working conditions to evaluate limits of human performance).

Time constraints for R&D activities are normally not a crucial issue, as in the rest of cases described in this paper, but interaction with other engineering teams for completing design is an important issue, in order to confirm findings and technical approach to consider for a design.

### 2.2 Case 2: Validations as part of Generic Licensing Activities

For Case 2, the methods and approaches applicable for a validation can be more limited for several reasons:

1. A technological approach has already been selected as part of a design, so the testing of different technological approaches in a validation at this stage seems meaningless
2. The necessary validations may depend on regulatory requirements and the existing level of detail of a design. For some countries a certified design may not need implementation and consequently no ISV applies, while in others the contrary can occur.

Because of above listed reasons, the possibilities of performing an ISV seem low, although not impossible. However other validation approaches that are not an ISV are possible, with important differences in the method to follow.

An example of a validation approach for this Case 2 is the validation of the function assignment, as proposed in IEC standards, and that does require in principle an integrated design (applicable early in the design stage). Another approach for Case 2 ISVs is performing partial validations with similar consideration as described in Case
1. However, for Case 2, the execution time constraints and interactions with other engineering teams can be more significant than for Case 1, as there is an official schedule to certify a design. Besides, the scenarios to follow need to be more realistic (in terms of duration, absence of interruptions and similarity of what would happen in a real design).

### 2.3 Case 3: Validations as part of Specific Design Licensing Activities

With regard to Case 3, two validations approaches can be considered:

1. Case 3a: Validations, or partial validations, associated to:
   a. Element 7 of NUREG-0711 (referenced as Tests and Evaluations during the HSI design process)
   b. Other validation approaches in ISO 11064 (in analysis and definition steps)
2. Case 3b: ISV, associated to:
   a. Element 10 of NUREG-0711
   b. Phase D in ISO 11064 (Step 10: Verify and validate detailed design proposal)
   c. IEC 60964/1771 (validation of the integrated control room system)

For Case 3a, the approach is partial validations for testing the HSI design being implemented, but still subject to changes as it is not totally finalized. Plant conditions in these partial validations can be role played by instructors or even better by candidate operators for the HSI. The latter has the advantage of having an early feedback of the HSI design by final users. Partial validations can be planned in a phased approach, with the HSI components that become available and the method greatly depends on the available simulator, facilities, operators, time and training.

Case 3b, is the genuine ISV testing, to be performed at the end of the detailed design phase and almost at the end of the HFE process for a specific site. As a consequence, ISV takes place when all the elements that form the integrated design are almost completely defined and developed. ISV can limit the wide variety of methods and measures to consider due to the important time constraints when executing a real implementation project, with important deadlines to fulfill and important interactions between multidisciplinary engineering teams, customer and regulatory bodies. The non-consideration of time constrains and interactions means to be out of touch with reality.

The above mentioned approaches, Case 3a and 3b, are consistent with an iterative process mentioned in international standards like ISO-11064 (that mentions that V&V must be integrated with the design process in a iterative process) or NUREG-0711 (that refers also to iterations in several Elements of this standard)

The rest of considerations to be described in this paper will be related to Case 3b of ISV, in which Tecnatom has more experience when licensing new designs for specific sites (as part of new plant or design modifications), and also because is the most relevant validation testing.

## 3 ISV for Specific Plant Design Licensing Activities

The following subsections will describe ISV considerations for Specific Plant Design Licensing Activities with regard to:

1. Applicability
2. Scope and Objectives
3. Validation Criteria/Issues
4. Performance Measures
5. Acceptance Criteria
6. Validation Scenarios
7. Test Design
8. Validation Results Analysis
9. Validation Reporting and Conclusions

### 3.1 Considerations Related to ISV Applicability

For new designs ISV is always applicable, but in the case of plant modifications guidance for determining ISV is not big, and the applicability may be determined depending on: 1) the type of modification under consideration (modification size, or importance of systems being modified), 2) the way plant operation is modified (if operator tasks are greatly modified) and 3) by considering the safety class of the systems involved in the design modification. The sometimes observed practice of considering ISV only in those cases where modifications are related to safety systems could be too limited, because non-validated modifications related to non-safety systems can change tasks, affect other HSIs or degrade operator performance.

### 3.2 Considerations Related to Scope and Objectives of an ISV

ISVs has to comply with nuclear regulations, consequently the objectives must be in compliance with the main standards related to this field such as (NUREG-0711, IEC-61771 or ISO-11064). ISV objectives should differentiate if it is for new plants or for design modifications, as the scope and technical approach methodology may not be the same. When defining objectives the following considerations arise:

1. The main objectives are always common to any ISV, being related crew performance for achieving safe plant operation. This is achieved by demonstrating an adequate crew performance with the design proposed/implemented. Other ISV objectives could be dependent of the HSI components to be assessed (considering new plant/design modifications).
2. Objectives should take into account that ISV for licensing a design is not a R&D activity; it is indeed in a schedule with other activities that affect each other and with a limited time frame for execution and reporting. A Nuclear Power Plant (NPP) under construction cannot wait until the end (when ISV takes place) to confirm the whole validity of the design or to test a technological approach.
3. When developing ISV objectives, HSI overall outcome objectives should not disregard the outcome of individual elements (each crew member, each HSI component). But ISV cannot evaluate everything with enough detail in a limited time frame, in a limited space with a limited number of observers. Consequently individual partial validations previously executed early in the design should have previously addressed individual elements in detail.
4. ISV objectives should consider that findings must be addressed/corrected, otherwise ISV could be meaningless. However ISV important recommendations that are produced at the very end of the HFE program can be very difficult to solve, especially if they have a great impact in the design, with the corresponding economical and schedule implications.
5. The facilities to be considered under ISV objectives should basically be the Main Control Room and Remote Shutdown Stations, where the most important plant operations are performed. Other facilities are also important, but the need of an ISV may not be so necessary, as other HFE activities such as verification could be enough.
6. Those activities/equipment with significant contribution to Core Damage Frequency (CDF) should be considered for ISV. The HRA identifies Risk Important Human Actions and these should be considered in scenarios for ISV.
7. Resolution of issues raised during the design process should be addressed prior to the execution of the ISV, unless the only way to resolve them is during ISV. ISV has many plant situations and HSIs to address and anything that is external to ISV and can be done independently should be scheduled separately.

8. Maintenance tasks in a NPP are almost infinite, being the majority external to the control rooms. The fact that they are normally performed in outages, do not have a critical time pressure for their execution and simulation is hardly available, draws to the conclusion that a separate HFE process, such as a verification during training activities could be enough from the HFE viewpoint. In any case, NUREG-0700 has limited support as is only for digital systems and consequently maintainability could be complemented with the use of other standards, such as EPRI-NP-4350.

### 3.3 Considerations Related to Validation Criteria/Issues to be tested in an ISV

The main objective in an ISV is to address crew performance by means of performance based scenarios. For this purpose it is useful to define "Issues to be Tested" or "Validation Criteria" that can later be used to substantiate or justify with more detail and in a standard way crew performance. The approach of Tecnatom is to consider:

1. Generic primary tasks performed by plant personnel (monitoring, situation assessment, response planning and response execution activities, based on a human performance model)
2. Secondary tasks (movements or navigation actions for accessing information)
3. Workload (for particular activities or within a whole scenarios, distinguishing cognitive and physical)
4. Situation awareness (for the global scenario)
5. Teamwork
6. Communication

The assessment and substantiation of these validation criteria has to be by means of performance measures/observations, as explained in the next section.

### 3.4 Considerations Related to Performance Measures in an ISV

For the execution of an ISV, the Team that observes crew performance has to mainly focus on the plant operation that each crew member will role play in specific scenarios to be defined and developed, relevant to the design under the scope of the validation.

There are many performance measures that can be considered for an ISV, among which the following can be considered: 1) time for performing actions, 2) precision, 3) accuracy, 4) frequency of actions, 5) amount achieved, 6) errors, 7) information exchanged (communication), 8) anthropometry/crew movements, 9) plant parameters values (along the scenario or at certain moments) or 10) crew ratings for specific questions.

These performance measures will be able to substantiate (in a qualitative and quantitative way) crew performance in the context of the scenarios role played. Performance measures can be taken during an ISV by direct observation, interviews and with various means of recordings:

- *Direct Observations*: ISV direct observations are the annotations that the observer's team will take during the scenario execution. Observations are subject to the method to be followed when executing the ISV, and of what is validated or observed. Observations may include directly performance measures or annotations by which performance measures can be calculated, being the latter the most common way to produce performance measures. As a consequence there can be cases where annotations may not produce any value for a given performance measure, as everything depends on the method followed, the interface being validated and plant conditions tested.

Direct observations knowing in advance what will happen in the plant and what crews should do seems the most efficient way to evaluate crew performance it terms of time needed (always short for an ISV) and the quality of the observations/performance measures. Consequently for a high quality ISV, observers´ training is indeed an issue. Considering the method followed by Tecnatom, a scenario log has to be prepared just after each scenario, documenting sequentially 1) who performed any actions, 2) each action performed, 3) when did the actions were performed (time), 4) what information was used, 5) which procedures were followed/used, 6) what information was exchanged and 7) HSI used. With this log performance measures can be derived or produced in many ways (frequency of actions, elapsed time for an activity, etc.).

- *Interviews*: Interviews after scenario execution have the purpose of asking crews to summarize what happened during the scenario and also formulating any specific question because of special issues observed during the scenario.
- *Video recordings*: Video recordings where scenarios were role played is another way to obtain observations and calculate performance measures. But the time necessary to observe video recordings and the limitations these can have (with regard of what recordings show in video and audio), advises this method as a supplementary method, for confirming direct observations or any doubt.
- *Simulator recordings*: Recordings of plant parameters during scenarios can provide an idea of the effectiveness of operator actions and plant performance.
- *Observer´s ratings*: These ratings can based on the direct observations of the Observer Team members during a role played scenario.

Many times the specification of a minimum set of observations and measurements that should be conducted, as part of the human factors validation of the MCR can be difficult, because it will really depend on what is really observed. Annotations may not provide values, and sometimes qualitative information to rate something as high, medium, low is the way out. Differences in values for the same measures sometimes can be significant and sometimes not, it depends on the measures and what is measured. Normally to achieve reasonably confidence for the same situation, similar values are expected for the same measures (e.g. number of tasks achieved), but sometimes no (e.g. number of times a displays is consulted looking for an information that exists in other displays).

### 3.5 Considerations Related to Acceptance Criteria in an ISV

Acceptance criteria to validate crew performance should be normally based on the comparison of previous designs. This comparison should be based on the experience obtained when facing plant situations or scenarios that are to be tested, being relevant of the design subject to ISV. The experience that comes from previous designs can be real experience (like normal operation experience), or simulator experience (like abnormal, transient and accident operation). From this point of view the benchmark approach seems the best approach:

1. Knowing in advance what crews should monitor, assess, plan and perform
2. Knowing in advance the expected medium time to perform all the actions

3. Knowing in advance the expected plant performance and final plant status

However, the benchmark approach can have its limitations, because sometimes no experience is available, like when using new technology never used previously (for the interface or as a new technological approach in the plant).

The Acceptance Criteria can be developed using as a basis the six Validation Criteria previously described. The six Acceptance Criteria should be written in a way to confirm adequate crew performance, as compared to previous designs and customized to the HSI under consideration. Standardized Acceptance Criteria would certainly help to assure the necessary confidence in the process, whatever method has been used. For design modifications that are very specific, Standardized Acceptance Criteria may not always be possible

The adequate performance in the Acceptance Criteria proposed should be substantiated defining and later observing performance requirements in the scenarios to be developed and that prescribe the activities to be role played by the crews during the validation. The scenarios must describe the expected plant performance and expected crew actions that will enable detection of any departure from the expected plant status and crew actions/performance. These scenarios are important because they will not only allow following operator actions, but also to understand what they are doing. Crew actions understanding will allow the observers team to judge what operators are doing and if the actions are correct or not.

The above described scenarios could take into account expected performance requirements that are the result of the design activities, the procedures developed as part of the design process, and risk important human actions that come from Human reliability Analysis.

### 3.6 Considerations Related to ISV Scenarios Descriptions in an ISV

Scenarios descriptions to be developed for an ISV can be of many types and should be representative for assuring reasonable confidence. For a new design they should cover all types of relevant plant conditions as those listed in NUREG-6393 or NUREG-0711, encompassing normal operations (including start-up and shutdown), malfunctions, transients and accidents. Normally the most risk important scenarios are what are more tested in new designs, because they are the most challenging situations for the plant safety and for crews. For design modifications, the scenarios should be relevant to the plant modification, covering the different plant situations in which that modification used (like system start-up, shutdown, maintenance, system failure or system monitoring in accidents).

The Observer´s Team that follows crew actions should be a multidisciplinary team with experience and knowledge in various fields, and not only in plant operation. Consequently scenarios to be role played should be developed in order to maximize the information included for appropriate guidance during scenario execution. The scenarios level of detail should not be excessive, taking into account that too much information could be useless, especially when multiples paths exist for crew actions (because plant contingency actions may depend on crew reaction time). A background in plant operation seems very necessary to understand what is happening in the plant and what crews do.

Scenarios descriptions could be like a script, describing what is going to happen so that observers can understand and anticipate what crews will do and what will be plant performance. These descriptions could include a brief overview for fast re-check and a detailed description in a time basis. The expected actions for each crew member should include the information they should detect, order, the procedures they should follow, as well as prescribed procedure actions they will follow.

The scenarios descriptions should be as realistic as possible, regardless if they are complex or not, because what counts is what crews will face in the real NPP. These descriptions can be developed from previous designs, where there is experience in plant situations, as commented in the benchmark approach for acceptance criteria. In the case of situations where experience is not available, scenarios could be developed from Task analysis or other preliminary tests performed (such as partial validations).

Apart from the relevant information in included in scenario descriptions, it is also important the relevant information to be collected by observers. Training of the type of information to be collected should be thorough, including: 1) time of the observations, 2) who made the actions, 3) the action done, 4) the HSI component operated/observed, 5) the information exchanged, 5) what movements were performed.

The information obtained from scenario execution could be summarized in a scenario log, including observations for the different members of the Observer Team. The different observations included in the log should be enough to know and follow what happened during the scenario and characterize the validation criteria for each of the scenarios role played. Later the validation criteria results for each scenario should be aggregated to provide general results for each scenario an later the validation criteria results for the whole set of scenarios (as described in section 3.8).

### 3.7 Considerations Related to ISV Test design

ISV Test design should consider the sufficient number of scenarios that can provide enough proof of the adequate HSI design and operator training to guarantee safe operation under all conditions relevant to the design. The scenarios can be short or long, or consider different plant operations in each of them (as described in section 3.6).

For test design, ISV should also consider the number of crews to participate and the sequencing of scenarios for each crew. NUREG standards (such as 6393) provide enough guidance about these two issues.

With regard to the number of crews to participate in an ISV, there are studies intended to analyze and provide the number of crews that are statistically significant. But simpler approaches are possible, like considering 3 to 4 different crews. In the case of 4 crews, this means that more than 50% of the user population has participated in the testing, taking into account that a plant unit normally has 7 crews (or more than 28% of the user population for a two unit plant).

With regard to scenario sequencing, Tecnatom experience shows that crews may transmit ISV experience and this can produce biased results. Consequently adequate scenario sequencing, so that crew will not know in advance the scenario they will face is a very important issue when planning the ISV.

Apart for the previously mentioned considerations for ISV test Design, it is very important to consider ISV duration as a licensing activity, and that should not be underestimated because many other interdependent teams depend on their results to finalize plant construction and licensing. Duration may depend if it is a new design or a design modification (shorter in many cases), and that determine the number of scenarios, and crews that will participate. ISV duration should not only consider the time of scenario execution, but the necessary time for documenting the results and providing conclusions. The latter can be up to three or four times the time necessary for scenario execution. The experience by Tecnatom indicates that as a minimum for new designs a one month duration (half working day) for each crew can be enough for covering the ISV scope, but this greatly depends on the type of plant, operational experience (considering the validation cases previously described), each scenario duration and specific regulatory requirements. Finally for an appropriate and reasonable ISV execution time, test design should not mix different types of validations (such as functional test).

Confirmation of ISV results for achieving and maintaining reasonable confidence in the ISV, can be achieved with:
1) The Human Performance Monitoring activity/strategy (as described in Element 12 of Nureg-0711), 2) Operational Experience Review, 3) Plant performance indicators, 4) Human Reliability Analysis, 5) and Training Evaluation.

### 3.8 Considerations related to ISV Validation Results Analysis

After the ISV execution, there are a series of documented role played scenario results that summarizes crew performance. A proposed approach when analyzing the obtained results is:

1. Step 1: Analyze individually each scenario, preferably documented in a log including also recordings and interviews materials. The collected information will allow identifying those special issues or events that are representative from a HFE viewpoint and obtain performance measures (numerical or qualitative).

   Performance measures and events will allow characterizing the validation criteria, and consequently rating the acceptance criteria based on the identified validation criteria. Departures from the expected and planned scenario response should be documented as Human Engineering Discrepancies (HEDs), for subsequent analysis.

2. Step 2: Compare the same scenario role played by each different crew, to detect and draw conclusions about common and different results/findings with regard to human and plant performance. This will allow to analyze human performance variability under the same circumstances (can be meaningful or not, depending of what is observed) and to characterize the validation criteria for the same scenario.

3. Step 3: Compare the results of the different scenarios to characterize the validation criteria for the same scenario for the facility under the scope of the validation and consequently draw general conclusions.

The HEDs that come for the ISV should have an analysis process to decide if design changes are necessary or not (to the procedures, to operator training, or to the HSI). The absence of changes should be justified, as well as the description of the changes to be performed when necessary (by the appropriate team). The need of an additional validation for checking how certain HEDs are solved would depend if they are important from safety point of view. An iterative validation process coming from the design would minimize the possibility of an additional validation that would have very important economical and schedule consequences. Normally HEDs would only need a last verification that could be continuous via the Human Performance Monitoring process.

### 3.9 Considerations related to ISV Reporting and Conclusions

With regard to the final validation report, Tecnatom perspective is that it should include:

1. Test and Evaluation conditions description
2. Summary of Results per HSI component, per scenario, per validations item/objective
3. ISV Acceptance Criteria fulfillment
4. ISV General Results and Recommendations
5. ISV Conclusions
6. ISV Report Appendices (including Scenario Logs, other scenario recordings, prepared HEDs)

A complete document should not only provide the conclusions of the overall ISV process, but also the minimum information that supports the conclusions provided. Conclusions should provide information of how the process was followed and how the gathered observations conclude how well the validation criteria and associated acceptance criteria are supported.

## 4 CONCLUSIONS

This paper has outlined the different types of validation approaches/contexts from the HFE point of view that can be applicable for NPPs designs. The different approaches are not mutually exclusive and can be complementary, depending on the context where the HFE activities take place.

Considerations related to ISV applicability, scope, objectives, validation criteria, acceptance criteria, performance measures, scenarios, test design, analysis of results, reporting and conclusions have been described from the Tecnatom experience. In any case there is no unique approach when defining a planning a ISV, as the context (R&D vs. Licensing), the type of plant involved, the scope (new or modified design), the customer, and applicable regulatory body requirements may condition in many ways how a ISV should be planned and executed (by an independent team).

The experience of Tecnatom applying HFE in modernizations and design modifications show a positive effect in improving the associated HSIs and acceptability of the final user, confirmed in the final ISV. However the fact that ISV is performed at the very end of a designs process makes necessary preliminary or partial validations for assuring that no major design changes are necessary at the end of any project, when it is very difficult.

The guidance provided in international standards include in many ways generic information of what should be considered when planning and executing a validation or ISV, but do not provide detail enough of what can better for planning and executing an ISVs. From this point of view this paper has described the perspective of Tecnatom for validations and ISV, in order to help identify approaches and methods that could be considered in the future for future guidance and standardization, for each of the contexts applicable.

**Paper for the WGHOF Experts' Workshop on Human Factors Validation 19th- 21st Feb, 2015.**

# What do we mean by reasonable confidence in human factors-related validations of integrated system performance for NPPs?

**Alice Salway**
Canadian Nuclear Safety Commission Alice.Salway@cnsc-ccsn.gc.ca

## 1 INTRODUCTION

There are two parts to the discussion of this topic:

1. What is meant by "reasonable confidence" in relation to validation
2. What it is that we have "reasonable confidence" about, concerning human factors validations

While this paper primarily considers integrated system validations of nuclear power plant control rooms, the discussion also applies to integrated system validations of system components that have been modified, both within and outside the control rooms. The main difference between the two types of validations is the smaller and narrower scope for the component validation. Integrated system validation can also be conducted beyond the control room or component applications, to include validation of minimum staff complement provisions, maintenance tasks, and emergency response tasks.

For brevity, this paper presents stereotypical profiles of various stakeholders. It is acknowledged that the real world is somewhat more subtle.

The views expressed in this paper are those of the author and do not necessarily reflect those of the Canadian Nuclear Safety Commission (CNSC).

## 2 DEFINITIONS

The BS EN ISO 9241-210 [1] definitions below provide context to this discussion paper:

**Validation** – confirmation, through the provision of objective evidence, that the requirements for a specific intended use of application have been fulfilled. NOTE: Validation is the set of activities ensuring and gaining confidence that a system is able to accomplish its intended use, goals and objectives (i.e., meet stakeholder requirements) in the intended operational environment.

**Verification** – confirmation, through the provision of objective evidence, that specific requirements have been fulfilled. NOTE: Verification is the set of activities that compares a system or system element against the required characteristics. This can include, but is not limited to, specified requirements, design description and the system itself.

**Usability** – extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency, satisfaction and context of use.

**Human-centered design** – approach to systems design and development that aims to make interactive systems more usable by focusing on the use of the system and applying human factors / ergonomics and usability knowledge and techniques. NOTE 1: The term "human centered design" is used rather than "user-centered design" in order to emphasize that this part of ISO 9241 also addresses impacts on a number of stakeholders, not just those typically considered as users. However, in practice, these terms are often used synonymously. NOTE 2: Usable systems can provide a number of benefits, including improved productivity, enhanced user well-being, avoidance of stress, increased accessibility and reduced risk of harm.

**Stakeholder** – individual or organization having a right, share, claim or interest in a system or in its possession of characteristics that meet their needs and expectations.

## 3 WHAT DO WE MEAN BY “REASONABLE CONFIDENCE IN VALIDATION”?

The term “reasonable confidence” is used infrequently at the Canadian Nuclear Safety Commission (CNSC), and it has not been defined by this organization. For the purpose of this workshop “reasonable confidence” is taken to have a generic meaning, similar to “reasonable assurance”, but *without* the regulatory implications of the term [2].

In practice, the meaning of “reasonable confidence with respect to human factors validations” is likely to depend on the specific stakeholders, the organizations they work for, their professional backgrounds, and their objectives for the validation. An informal poll of what “reasonable confidence in validation” could mean, generated the following list:

- High confidence in a low likelihood of failure, in relation to the system goals and requirements
- Confirmation that “the right system has been built” to achieve the system goals and safety objectives
- The validation has been successful in capturing weaknesses in the integrated system design and support provisions, and it is now considered acceptable for operation
- Validation results are obtained to demonstrate that system performance—including the humans in the system—meets the specified performance requirements or criteria
- Engineering or expert judgment that the validation results are acceptable
- A level of confidence based on facts, testing, knowledge, experience, and/or observations, which is deemed to be adequate
- Assurance that the equipment or system will perform as intended by the designers
- Assurance that the system or equipment supports the assumptions and analysis about human actions made in the plant’s safety report
- Demonstration that the specific roles of the system users, with specific equipment, procedures and training, can carry out a task (or set of tasks), in a specific environment, to the required standards of performance, effectiveness and safety
- Evidence that the plant is more robust than needed for normal use alone, to allow for emergency situations, expected misuse, or degradation
- Tasks that use or apply to safety-related and safety-significant equipment can be carried out to the required levels of performance
- Tasks relating to the full plant design envelope (including design extension conditions) can be carried out to the required levels of performance
- The validation measurement approach has acceptable certainty, related to the way that the performance measurement has been carried out
- An acceptable margin of error is associated with the sample of validation measurements. commonly, the likelihood that the results from the sample reflect the true results is 95%

- The appropriate significance level has been applied, related to hypothesis testing, i.e. the probability of incorrectly rejecting the hypothesis being tested, usually between 0.01 to 0.1

These interpretations fall in two groups. The largest number relate to a practical, empirical approach for declaring that a system is acceptable, with a few that relate to statistical methods.

Integrated system validation differs from usability testing in several ways. In usability testing it is possible (and often desirable) to have narrow and well-defined questions of interest. The trials can be controlled and focused on a small number of variables. Despite the “quick and dirty approaches” and iteration, usability testing is more amenable to valid use of statistical methods. However they may not always be used, dependent on the formality and the purpose of the usability tests.

Integrated system validation is less amenable to the valid application of statistical methods. Bringing together the system elements in an integrated system validation results in high complexity, through integrating the technical systems, support systems (such as training, procedures), and the people in the context of the plant organization and management systems. When considered in this way, the behavior of the whole system is like an organic system due to the complex interactions, feedback mechanisms and variability in its elements and outputs. Integrated system validations have many questions of interest and less certainty and control of variables than in usability testing. Numerous variables and variation are introduced when bringing together the many facets important to human performance in high-fidelity simulations. Integrated system validations are resource intensive in terms of simulator time, trained operators and other relevant roles, and the people needed to stage the validation such as role players, and those involved in data collection. Large numbers of trials cannot be carried out in integrated system validations, due to the high complexity, multiple questions of interest and high resource demands, which relate to the real-world constraints of time and money. Usually there are low numbers of “trials” and low numbers of “subjects”, which present difficulties for valid experimental design and statistical analysis.

## 4 WHAT WE HAVE “REASONABLE CONFIDENCE” ABOUT

To address this point, the different stakeholders and their perspectives of human factors-related validations need to be considered.

The three main stakeholder organizations are:

1. System Vendor
2. System Customer
3. Nuclear Regulator

For the Vendor and Customer, the human factors validation activities and results are often viewed as a formal system acceptance test related to contractual requirements, although from different perspectives. Regulatory requirements and expectations are usually included in the contractual requirements, given that the Customer must obtain and comply with licenses in relation to the lifecycle of the nuclear facility.

The key stakeholder perspectives with regard to human factors-related validation activities could be considered as follows:

- The Vendor uses the validation activities as part of the demonstration that the human factors-related system requirements have been met, from a contractual perspective. If problems are identified, the vendor will use the validation data to correct the problems, so it is useful if the data collected have diagnostic value, as well as demonstrating conformance with requirements. Prior to the validation, the vendor organization should already be confident that the system will perform as anticipated, because the part-system testing with representative end user roles during the system development will have indicated that the system design is usable and that it meets requirements. In addition, the validation results contribute to the Vendor's reputation and track record, as does the operational performance of the system. A generic design, which aims to conform to a range of national regulatory frameworks, may be validated by the Vendor, the results of which will provide early assurances to Customers and Regulators that licensing is feasible and that the design is in conformance with high-level safety and performance goals.
- The Customer uses validation activities to obtain assurance that the formal system requirements have been fulfilled, but also to provide practical assurance that the system will perform as intended and desired in the specific plant environment with the specific workers, infrastructure and system support provisions. In other words, the Customer uses validation activities to determine if it is satisfied with the system that it is procuring. Appropriate shaping of validation activities and consideration of the results is a "smart buyer" characteristic. Meeting the regulatory requirements is also a key consideration. Shortcomings need to be identified so that they can be remedied, from the perspective of ongoing system performance, costs, safety, and regulatory considerations.
- The Regulator uses validation activities to confirm that the people in the system have been appropriately considered in the design with regard to regulatory requirements and expectations to protect health, safety, security and the environment. The validation results should demonstrate that it is feasible for workers to perform their tasks as they relate to the regulatory goals for the facility. Successful performance during the integrated system validation also provides assurances that the wider system support issues have been analyzed, planned and implemented appropriately, e.g. staffing, training, procedures, security, provision of tools, PPE, spares, consumables, etc.. Final integrated system validation represents a "last chance" activity to identify emergent issues and to provide assurance that the system fulfils its requirements and it will be safe to operate.

Given these views, the system Vendor and Customer will both use the validation results to provide reasonable confidence to confirm that contractual requirements for the system acquisition and acceptance have been met. The Customer will also be a licensee, which therefore has considerations associated with its responsibility for meeting regulatory requirements and expectations, where these have not been covered in the contractual requirements. The validation therefore provides reasonable confidence to the Customer/Licensee that the regulatory requirements have been met. The Customer/Licensee may also have considerations for processes and performance related to its business, which are not part of the regulatory framework.

The Regulator generally reviews and observes a sample of the validation activities, to have reasonable confidence that the regulatory requirements have been met. The Licensee is directly responsible for the health, safety, security and the environment at the plant, while the Regulator exercises oversight that the regulatory requirements have been met and that compliance is maintained. The Licensee will have more information than the Regulator to have confidence that the regulatory requirements have been met, because the Regulator usually considers only a sample of the available information. Based on the relative information available to these two stakeholder organizations, the Customer /Licensee should have greater confidence that the regulatory requirements have been met than the Regulator, because its' information will cover all validation considerations, not just the ones sampled by the Regulator.

## 5 DISCUSSION

From the perspectives of the human factors specialists working for all three stakeholder organizations, in addition to meeting formal requirements, validation is a confirmation that the system users are supported to carry out their tasks by the whole system, including the organization, management and user support provisions. Although human factors specialists may not be responsible for developing these wider provisions, an effective system design considers and influences them to optimize performance. Various factors are pertinent to human-system performance, such as the number of available workers to carry out tasks, their knowledge and skills, the procedures provided, the organization of work including supervision, provision of necessary equipment such as tools, personal protective equipment (PPE), and the physical environment where the work is carried out. This perspective of "supporting things that go right" through human factors in design is certainly not new, although it aligns with the "Safety II" approaches that have emerged in recent years. [3] [4] [5] [6].

"Supporting things that go right" relates to the approaches for selecting scenarios and tasks to be included in the validation. While reasonable confidence can be associated with a graded approach to risk, it should be remembered that risk is a product of probability and consequences. There seems to be a tendency for validations to focus on Design Basis Accidents, although analysis and experience have determined that a plant will spend a mere fraction of its operating life in this state. Although it is clearly important to consider accident scenarios and their mitigation due to the high consequences of failure, scenarios that would have high probability but lower consequence outcomes should be considered as well. It follows that "supporting things that go right" in Normal Operations and Anticipated Operational Occurrences should also receive attention during validations.

System requirements [7] are important to the various stakeholders' perspectives. If the system is validated against requirements, these need to be formally developed and specified during design, to include human-focused requirements for human factors-related validations. Requirements specification, verification, validation and requirements management, as well as human factors in design are standard components of systems engineering approaches [8] [9] [10] [11].

In Canada, there is regulatory emphasis on the process of human factors in design for new nuclear power plants and for modifications [12] [13] [14] [15], as well as on the verification and validation results themselves. This approach considers a variety of information, which ideally converges to indicate and predict that the system supports the workers within it to achieve the system performance and safety goals. Indications of inadequately resolved system performance issues or safety problems (which can relate to technical, human or organizational aspects), or non-conformances in the human factors in design process serve to initiate deeper regulatory scrutiny and oversight.

For a non-power reactor design project where it was not feasible to conduct an integrated system validation until after the system implementation, CNSC staff accepted a case based on converging evidence from the design development activities, which indicated that the system was likely to meet human factors requirements related to safety and performance. On this basis, the system was commissioned before conducting validation activities because the regulatory and safety risks were identified as low by both the Licensee and the Regulator. Similarly, current approaches for the design of military systems rely on systems engineering approaches, which use iterative and integrated multi-disciplinary design development approaches, including testing, modelling and simulation. In the military domain, the emphasis on a large integrated system validations at the end of the design phase has steadily reduced over the last fifteen years, whereas ongoing post-implementation monitoring of system process and performance issues – including human factors and human performance – is more heavily emphasized than before. This change may be due, in part, to the now widespread use and refinement of system engineering approaches to encompass the complexity of contemporary military systems.

Even with considerable development testing with representative user roles, there are many factors that can provide surprises with respect to human performance when the entire system is assembled and operated. Validation enables some of these emergent problems to be identified and fixed, and it is inevitable that some sort of human-related emergent problems will be discovered, often where factors that influence human performance have been inadvertently mismatched. Unanticipated issues often continue to emerge in the early period of system use, which is why post-implementation monitoring is of practical value. If there is no formal identification and fixing of human factors-related problems, the system stakeholders just have to live with them. This can result in a system that is not actually being used as the designers intended, which may not be efficient or effective and may even have serious safety implications.

# 6 CONCLUSIONS

The issue of "reasonable confidence" in integrated system performance largely relates to practical demonstrations to indicate that system and safety goals will be met, including the contractual and regulatory requirements.

The different perspectives of the various stakeholder organizations need to be considered when discussing validations activities among diverse participants. All perspectives have some common ground, but there is a very different emphasis across issues and outcomes. The varying perspectives must be clearly stated, otherwise the three stakeholder groups could unknowingly be at cross-purposes when discussing validation. This issue should be considered throughout the NEA/WGHOF workshop discussions of validation. A further stakeholder group of "researchers" will add to the diversity of perspectives represented in the workshop discussions.

It will be important to focus on the common issues, and to be aware of fundamentally different views and perspectives. A common issue for system Vendors, Customers and Regulators is safety, but taking slightly different views. A validation focus on "supporting things to go right", aligns with contemporary "Safety II" approaches, and may be meaningful to all the stakeholder groups in relation to both plant safety and performance.

For the smaller scope of component modifications it is feasible that a system acceptance approach could consider the approach, content and implementation of a specific design's human factors in design process, the converging results of relevant design development activities (such as task analysis, stakeholder inputs to the design, operating experience reviews, modeling and simulation, safety analyses, usability testing and expert reviews), and also the activities and outputs of system support activities during the design, i.e., development of training and procedures. The extent of integration of the human factors in design work with that of the wider design project is also a relevant consideration, because human factors in design is minimally effective when conducted in silos. This information would be collated and analyzed to present a case for reasonable confidence that the system will perform safely, and as desired and anticipated. If this approach is taken, active post-implementation monitoring is necessary. The aggregate of the changes being made to the plant also need to be considered when deciding the suitability of this approach, where there are flaws in considering several separate modifications in the same or related systems with multiple and separate applications of this approach.

## 7 REFERENCES

1. International Organization for Standardization, *Ergonomics of human-system interaction, Part 210: Human-centered design for interactive systems*, BS EN ISO 9241-210, ISO, Geneva, Switzerland (2010).
2. Electric Power Research Institute, *Program on Technology Innovation: 10CFR50.69 Implementation Guidance for Treatment of Structures, Systems, and Components*, Report # 1011234. EPRI, Palo Alto, USA (2006)
3. E. Hollnagel, *The ETTO principle: Efficiency-thoroughness trade-off. Why things that go right sometimes go wrong.* Ashgate, Farnham, UK (2009)
4. S. Dekker, *The Field Guide to Understanding Human Error*. Ashgate, Burlington, USA (2006)
5. S. Dekker, *The Field Guide to Understanding Human Error*. Ashgate, Burlington, USA (2006)
6. S. Dekker, *Just Culture: balancing safety and accountability*. Ashgate, Farnham, UK (2007)
7. The Institute of Electrical and Electronics Engineers, *IEEE Guide for Developing System Requirements Specifications*, IEEE Std 1233, IEEE New York, USA (1998)
8. International Council on Systems Engineering, *Systems Engineering Handbook: A guide for system life cycle processes and activities.* INCOSE, San Diego, USA (2010).
9. National Aeronautics and Space Administration, *Systems Engineering Handbook, NASA/SP-2007- 6105 Rev.1.* NASA, Washington, USA (2007)
10. The MITRE Corporation. *Systems Engineering Guide*, MITRE, Bedford, USA (2014)
11. International Organization for Standardization, *Systems and software engineering – System life cycle processes*. ISO/IEC 15288 IEEE Std 15288-2008, ISO, Geneva, Switzerland (2008)
12. Canadian Nuclear Safety Commission, G-276, *Human Factors Engineering Program Plans*, CNSC, Ottawa, Canada (2003)
13. Canadian Nuclear Safety Commission, G-278, Human Factors Verification and Validation Plans, CNSC, Ottawa, Canada (2003)
14. Canadian Nuclear Safety Commission, Physical Design: Design of Reactor Facilities: Nuclear Power Plants, REGDOC-2.5.2 CNSC, Ottawa, Canada (2014)
15.
16. Canadian Standards Agency, Human factors in design for nuclear power plants, CSA N290.12-14 CSA, Toronto, Canada (2014)

## ISV – Reduction of Requirements, Clear and Simple Measures, Hypotheses and Success Criteria

### Reduction of Requirements

Clearly, requirements from national and international standards are an important source for the ISV – be it methods requirements (how to conduct an ISV, how to select scenarios, what to measure etc.) or design requirements (what to validate).

However, in the end, ISV aims to empirically demonstrate that the integrated control-room system (at the minimum the control room(s) environment, main HSI to monitor and operate the plant, supplemental for additional tasks like communication with staff outside the control room, operating procedures and the shift team at an intended training level) meets the intended purpose, i.e. that the control room staffing is able to effectively (and efficiently) and thus safely operate the plant.

Parts of the design requirements are rather global (in line with the general goals of ISV as described above), even vague, and (would) have to be brought to a level of details suitable for the empirical setting, which leaves lots of freedom in how to do so, and can thus be questioned or even doubted by different stakeholders of the project. On the other hand, standards requirements often enough list very detailed and numerous design requirements that a suitable for validation – one by one. However, due to their shere amount, validation of all of these requirements during ISV is impossible, from practical but also methodological point of view. This can be easily understood when one thinks of experimental settings (which the ISV is not, however, the same is true to certain degree for other empirical settings like quasi-experiments): The probability of a statistical hypothesis being refused although valid, or is kept although not valid (error of 1st kind or 2nd kind) increases the more hypotheses are set up for on an experiment. Finally, there are detailed requirements available, which are nevertheless formulated in ways not (directly) suitable for empirical validation (example would be that the HSI of conventional panels have to be optimized for the intended purpose – the question would be how to validate that an HSI is „optimal“.)

As a result, ISV should focus on the very aim of it as described above (or described similarly), and thus should derive testable requirements from that aim, with a strong and unambiguous relationship to the aim.

These requirements mostly can be found in the area of operator / shift team performance. The two dimensions of performance are the effectiveness (the degree to which the single operator or the shift team as a whole can successfully perform their tasks / can reach certain performance goals), and the efficiency (the time the operator / the shift team needs to reach these goals). Note these dimensions of performance are in fact dimensions of usability in terms of ISO 9241-11, although ISO 9241-11 defines efficiency more in terms of amount of resources to reach the goals, and identifies further on in 924-110 the influencing factors for efficiency (learnability, self-descriptiveness, etc.).

In fact, shift team performance is already a strong reduction of the aim of ISV (the integrated-control room system is to be validated, parts of which the shift team is). Ratio behind is that the shift team performance is the integrated results of all elements of the integrated control-room system, and with one, system’ malfunctioning, or interfering with another system, the resulting overall performance is low as a consequence.

Those performance measures (effectiveness, efficiency) cannot be generally assessed, as they depend on the very situation of each scenario selected (in terms of usability: usability is context- sensitive: an HSI can be usable for one task, or one user type, while poor for the another). This requires thus careful definition of performance criteria both for effectiveness and efficiency, for each ISV scenario selected. The good message about these measures and particularly the success criteria is that they do not require – nor allow (!) – statistical evidence. Reason is that – given the success criteria are defined in a balanced and valid way, i.e. criteria must not be trivial on the one hand, but most also be too strict or numerous – all of these criteria are to be met. This means, that every shift team (for shift team performance criteria), and every operator (for operator performance criteria) have to reach the criteria. Reason is that, if the performance criteria are validly related to the purpose of ISV, i.e. that the plant can be (always) safely operated by the shift teams, one failure is one evidence of unsafe operation and thus not acceptable.

This might be a strong and hard-to-reach evidence. On the other hand, one failure instance during ISV does not mean a failure of the integrated control-room system as such, but only for the context (and, of course, for certain areas the scenario was supposed to be representative for, e.g. severe accident operation when performed by display-based HMI in combination with paper-based procedures). Thus, re-design would be indicated only for those areas of generalization.

However, efficiency and effectiveness are not independent variables, and the relationship between should be clarified: Inefficient performance – when measured in terms of time as suggested – probably will lead to ineffective performance, because a goal that ‚is reached‘ too late, is in fact not reached. (an ‚effective‘ manual containment isolation is, when performed too late, not successful..). It could be that efficiency is an additional measure relevant for only time-critical situations, while the effectiveness is always a valid measure. However, the term ‚time-critical‘ is rather vague, which can easily be demonstrated when asking different operations experts for time-critical situations – the answers might differ. So, better than talking about time-critical situations, is to define time requirements coming from mostly safety and risk analysis.

In-between the main ISV measures as defined above, and supplemental measures as indicated in the next section, might be the human errors: Here, the relationship between task performance and the concept of human errors should be clearly defined, respectively clarified: One could argue that the total absence of human errors (depending of course on the definition of human errors) task performance mandatorily or naturally is acceptable, or even is a result of the absence of human errors. On the other hand, it is possible that task performance is acceptable even if human errors occur. It has thus to be clarified whether human errors can be used as indication for task performance, or should be seen as additional measure in ISV.

### Supplemental Measures

When task performance could be successfully validated, i.e. no violation of effectiveness nor efficiency could be witnessed for any of the scenarios selected during ISV, the integrated control- room system can be seen as validated.

However, supplemental measures provide further understanding (and further evidence, particularly if performance is near to the boundaries and criteria of acceptance) for the reason of good or lacking task performance. These are thus impacting factors of overall task performance.

Example for those measures are the construct of Situation Awareness: Situation awareness that is below a certain (but probably unknown) can result in wrong decisions (human error), leading to ineffective operation. Here the question is if there are acceptance criteria for situation awareness in special, and supplemental measures in general: Is it possible to define unacceptable levels of situation awareness – even though it does not result in task performance below acceptance criteria? The same is true for workload: is there a workload level that is at an unacceptable level even though the performance is acceptable? This is theoretically unclear; practically, it would require – as workload mostly is assessed via subjective quantitative measures – statistical evidence.

# ISV – Human Performance Evaluation: Needs, Considerations and constraints, Criteria, and Rationale

Jun Su Ha (junsu.ha@kustar.ac.ae)
Department of Nuclear Engineering, Khalifa University of Science, Technology & Research, UAE

Poong Hyun Seong (phseong@kaist.ac.kr)
Department of Nuclear and Quantum Engineering, Korea Advanced Institute of Science and Technology (KAIST), Republic of Korea

## Human Performance Evaluation

**Needs:** The objective of the ISV is to provide evidence that the integrated system adequately supports plant personnel in the safe operation of the relevant NPP [1]. The safety of a NPP is a concept which is not directly observed but must be inferred from available evidence. The evidence can be obtained through a series of performance-based tests. Consequently, if the integrated system is assured to be operated within acceptable performance ranges, the integrated system is considered to supports plant personnel in the safe operation. The operator's tasks are generally performed through a series of cognitive activities such as monitoring the environment, detecting data or information, understanding and assessing the situation, diagnosing the symptoms, decision-making, planning responses, and implementing the responses [2]. Hence, the HMI (Human-Machine Interface) design of a MCR (Main Control Room) should have capability to support the operators in performing these cognitive activities by providing sufficient and timely data and information in an appropriate format. Effective means for the system control should be provided in an integrated manner as well. If the HMI design has the capability, the operators can effectively monitor and detect data and information representing the plant status, understand the state of the plant system correctly, which also support appropriate diagnosing the plant system, decision-making, and thus responses planning, and then implement the responses. Consequently, the suitability of the HMI design of a MCR is validated by evaluating human (operator) performance resulting from the series of cognitive activities.

**Considerations and Constraints:** Attentions should be paid to some considerations and constraints, when human performance measures are to be selected and/or developed for the ISV. Firstly, considerations should be paid to changed operating environment. Especially the conventional analog based HMI is changed into digitalized one in Advanced Control Rooms (ACRs). As O'Hara and Robert pointed out [3], there are three important trends in the evolution of ACRs such as increased automation, development of compact and computer-based workstations, and development of intelligent operator aids. Increases in automation result in a shift of operator's roles from a manual controller to a supervisor or a decision-maker. The role change is typically viewed as positive from a reliability standpoint since unpredictable human actions can be removed or reduced. Thus the operator can better concentrate on supervising the overall performance and safety of the system by automating routine, tedious, physically demanding, or difficult tasks. However inappropriate allocation of functions between automated systems and the operator may results in adverse consequences such as poor task performance, out-of-loop control coupled with poor situation awareness, and so on [1]. In addition, the shift in the operator's role may lead to a shift from high physical to high cognitive workload, even though the overall workload can be reduced. Computer– based workstation of ACRs, which has much flexibility offered by software–driven interface such as various display formats (e.g., lists, tables, flow charts, graphs, etc.) and diverse soft-controls (e.g., touch screen, mice, joy sticks, etc.), is thought to affect the operator performance as well. Information is typically presented in pre-processed or integrated forms rather than raw data of parameters and much information is condensed in a small screen. The operator has to manage the display in order to obtain data and information which he or she wants to check. Hence poorly designed displays may mislead and/or confuse the operator and thus increase excessively cognitive workload, which can lead to human errors. Due to these changes of the operating environment, the operator's tasks in ACRs might be conducted in a different way from the conventional one. Hence enhanced attention should be paid to operator task performance and cognitive measures such as situation awareness and workload. Secondly, the evaluation of human performance should be practical and cost-effective (or economic). Since the aim of the performance evaluation is eventually to provide an effective tool for the ISV, evaluation techniques should be practically able to provide technical basis in order to get the operation license. In addition, the ISV is performed through a series of tests which require considerable resources (e.g., time, labor, or money) from preparation to conclusion. Hence cost-effective (or economic) methods which are able to save resources are required. For the best effectiveness all human performance measures selected and/or developed for the ISV should be evaluated simultaneously without interfering with each other. For example, if simulator-freezing techniques such as SAGAT (situation awareness global assessment technique) or SACRI (situation awareness control room inventory) are adopted for the evaluation of situation awareness, it is thought that the simultaneous evaluation of workload might be interfered by that of situation awareness (that is, the simultaneous evaluation of workload might be contaminated by the frequent freezing, which might require another same simulation of a test scenario for the workload evaluation). Thirdly, evaluation criteria for the performance measures should be clear. If it is not applicable to provide clear criteria, the criteria should, at least, be reasonable in the state-of-the-art. Human performance measures represent only extent of performance. The acceptability of performance in each of measures should be evaluated on the basis of performance criteria. There have been developed lots of human performance measures. Each of them has advantages and disadvantages (or limitations). Hence, multiple complementary measures should be used to complement each other

### Human Performance Criteria

Approaches to establishing criteria vary based on types of comparisons such as requirement referenced, benchmark referenced, normative referenced, and expert-judgment referenced [1]. Firstly, the requirement referenced is a comparison of the performance in the integrated system considered with an accepted and quantified performance requirement based on engineering analysis, technical specification, operating procedures, safety analysis reports, and/or design documents. Specific values in the plant parameters required by technical specification and time requirements for critical operator actions can be used as criteria for the requirement referenced comparison. When the requirement referenced comparison is not applicable, the other approaches are typically employed. Secondly, the benchmark referenced is a comparison of the performance in the integrated system considered with that of a benchmark system which is predefined as acceptable under the same or equivalent conditions. Thirdly, the normative referenced comparison is based on norms established for performance measures through its use in many system evaluations. The performance in the integrated system considered is compared to the norms established under the same or equivalent conditions. Finally, the expert judgment referenced comparison is based on the criteria established through the judgment of subject matter experts (SMEs).

### Rationale for Selecting Human Performance Measures

The following questions should be asked when adopting and/or developing human performance measures for the ISV:

- Q-1: Do human performance measures adopted and/or developed encompass the changed operating environments?
- Q-2: Are they practical and cost-effective?
- Q-3: Do they have clear performance criteria?

If all the questions are satisfied, the next question would be how to design test scenarios and integrate the evaluation results from multiple measures to conclude whether the integrated system is acceptable or not.

### References

[1] J. M. O'Hara, W. F. Stubler, J. C. Higgins, and W. S. Brown, Integrated System Validation: Methodology and Review Criteria 1997, NUREG/CR-6393, US NRC.

[2] M. Barriere, D. Bley, S. Cooper, J. Forester, A. Kolaczkowski, W. Luckas, G. Parry, A. Ramey-smith, C. Thompson, D. Whitehead, and J. Wreathall, Technical Basis and Implementation Guidelines for a Technique for Human Event Analysis (ATHEANA) 2000, Rev.01, NUREG-1624, US NRC.

[3] J. M.O'Hara and R. E. Hall, "Advanced control rooms and crew performance issues: Implications for human reliability," IEEE Trans. Nucl. Sci., vol. 39, no. 4, pp. 919–923, Aug. 1992.

## Envy in V&V:

## Preliminary Thoughts in Support of the WGHOF Workshop on Verification and Validation

Ronald Laurids Boring, PhD Idaho
National Laboratory

### *Various Vs*

I've given this short opinion piece on verification and validation (V&V) a somewhat cutesy and alliterate title, but I wish to make a few serious arguments here. Let me begin with the word *envy*. Envy here has two senses: (1) we are seeking after something that is not ours, and (2) we should emulate the desirable qualities of something else. The former is famously negative (e.g., Bertrand Russel's "One of the most potent causes of unhappiness is envy"), while the latter presents the opportunity for positive outcomes (e.g., Bertrand Russell's "When envy is inevitable, it should be used as a stimulus"). What does it mean to suggest that V&V is not ours—as in the human factors community's—to call our own? Further, what does it mean to suggest we should strive for V&V? I'll revisit those two points at the end of this paper. First, I wish to highlight two areas where I believe V&V has stumbled in the nuclear community: *when* it is performed and *what* measures are used.

### *Better Late Than Never?*

It is accepted in the human factors community that it is better to be involved early in the design of a system rather than later. This stems from the best window in the design cycle for our field to affect change. Change early in the design cycle—in the formative stages of system design—allows for the incorporation of user input to improve the design. Conversely, performing an assessment of a design late in the design stage—at the summative stage—risks finding fault in a nearly deployed system. Late-stage V&V hardly endears us as contributors to the end product, nor does it allow adequate time to fix issues that may surface in the system.

My key point here is that in the nuclear community, with our strong emphasis on summative evaluation in the form of integrated system validation, we potentially put ourselves in the position of doing human factors at the tail end of the design process, when we are relatively speaking least able to improve the design. There is nothing prescribing this tendency toward late-stage evaluation. It may be a simple confusion of guidance such as *NUREG-0700*, which is foremost a document guiding regulatory review at the completion of the design cycle rather than an exhaustive best practice for human factors. The propensity for late-stage V&V may also be a result of a certain disclosure hesitancy between the licensee and the regulator, in which the intermediate steps of a design—the designs with shortcomings that might be revealed through operator studies— are not readily shared as part of a license submission. The problem is that when V&V is relegated to a tail-end activity, we have not necessarily engaged in a process of system improvement based on user input and evaluation. Nor have we documented lessons learned in the design process. We tend to focus on demonstrating that the overall system as designed actually worked. We haven't demonstrated that the design evolved to the point of working. We seek to rubber stamp design rather than actively refine it.

It is my opinion that we as a human factors community need to reassert V&V not just as integrated system validation but as part of an iterative user centered design process. Experience in other domains (e.g., educational testing, safety cases, and quality control) reveals the advantages of early and frequent sampling of progress to demonstrate a successful process. We need to understand and document stumbling blocks that weren't good design ideas. These ideas need to be shared by licensees as welcome byproducts of the design process. Equally importantly, design foibles that are overcome through early- stage and iterative V&V should be championed by regulators as artifacts of an effective human factors process.

### *Measures That Don't Measure Up*

The use of the performance measures in V&V is sometimes driven by the state of the art in human factors, not by their practical utility. Please do not misinterpret this statement to be a criticism of the many solid human factors approaches we see represented in the literature and at this workshop. There is a need for better measures, whether to refine existing measures or develop new ones. But, the fundamental question remains: *Are we actually measuring what we need to in order to perform the V&V?*

At a superficial level, the purpose of V&V is to establish that operator performance while using a system meets a minimum standard. That minimum standard may be set in terms of safety, reliability, workload, or other measures. The challenge is that these standards— and how to measure them—are not always clear. We need to do more work to establish the expectations of acceptable performance so that V&V studies can benchmark to that level. Without such clear standards, we risk the distractions of measurement novelties. Situation awareness, eye tracking, and physiological measures—while certainly constructively pushing the bounds of psychological measurement—may prove to be surrogates for the measures we actually need for operator performance. Again, I do not wish to criticize research that uses these types of measures, which may in fact be the key to understanding operator performance better. My point is that these and any number of exotic advances in psychological measurement do not necessarily help us perform V&V better than we currently do. We must stop, catch our breath for a moment, and determine how different measurement tools available to us as researchers and practitioners actually help us understand operator performance. If our measures do not specifically verify or validate, we should discard or refine them. We must not be distracted by a gluttony of measurement.

### *Reprieve from Envy*

The term *physics envy* has been suggested to describe the desire of so-called softer sciences to achieve the mathematical precision and law like certainty of physics as a hard science. It is a variant of this physics envy that I believe underlies the two issues I've highlighted here.

First, V&V researchers strive for a sort of finality in the findings of V&V. There is a desired precision and conclusiveness in saying a system has been verified and validated. It suggests that there's no room for error or refinement. The book is closed once the V&V is performed, and there's no need for questions. Perhaps this is one root of the tendency for late-stage V&V over early-stage efforts. This is akin to physics envy in the sense of striving for law like precision in our findings. Rarely are the findings from V&V so conclusive, even when they must stand up to regulatory scrutiny. That's okay. Humans are remarkably resilient to consistency and classification. We must make our peace with the imprecision of V&V. A better approach is to show the trajectory of the findings. This is demonstrated through iterative evaluations early in the design—showing the refinement of the system design and the improvement of operator performance while using the system. It is the process of improving the design—not the immutability of the V&V findings—that determines the system is successful and usable by operators.

Second, as noted, V&V researchers have tended to gather increasingly complex measures of performance. It might be argued that this is in pursuit of a more scientific and scrutable set of findings rather than the subjective measures we must often employ in our studies. Certainly, the pursuit of better measures should be applauded. But, these measures must not be applied simply to further the hope of greater scientific precision. A good measure is any measure that provides insights into operator performance. It is not simply the quality of the measure but rather the quality of clearly matching the measure to V&V objectives that will ultimately prevail the science of V&V.

As V&V researchers and practitioners, we sometimes envy fields that provide highly conclusive findings, and we compensate with an ever increasing arsenal of measurement methods. These forms of V&V must stop! Instead, we should embrace the evolving nature of the findings afforded by early-stage evaluation using relevant measures to support our analysis. We have no reason for envy. We have every reason to refine our processes and measures to best reflect operator performance and system interfaces. In short, we have every reason to continue verifying and validating V&V. I would argue we are at the early stage of this process and can still affect considerable improvement.

### *Disclaimer*

The opinions expressed in this paper are entirely those of the author and do not represent official position. This work of authorship was prepared as an account of work sponsored by Idaho National Laboratory, an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately-owned rights. Idaho National Laboratory is a multi-program laboratory operated by Battelle Energy Alliance LLC, for the United States Department of Energy under Contract DE-AC07-05ID14517.

## Test Scenarios and Test Design

Kenji Mashio Mitsubishi Heavy Industries, Ltd.

Test scenario and test design development is a key factor for a success of the V&V.

### 1. Sampling dimension

When we evaluate a new HSI system, it is crucial to choose representative operational conditions and significant human actions with possible worst situations (e.g., with stress) (so called "sampling dimension" in NUREG-0711, rev.3) since it is impractical even unnecessary to implement all possible operational conditions, operator's actions/tasks, and all situations.

When we define sampling dimension, as NUREG-0711, rev.3 guides, it is important to ensure followings;

i) Representative operational conditions cover sufficient range of plant mode, normal (including start-up, shutdown, and refueling), transient and accidents. When digital I&C and HSI platform is introduced, failure of DCS causes to degrade HSI usage so that this condition should also be considered.

ii) Operator's action include design assumed human actions, important human actions (HAs) from PRA/HRA, credit manual actions in safety analysis, and manual initiation of protective actions, and the other design assumed HAs (e.g., time critical HAs and high-workload HAs from TA with comparison with design assumptions. Note that automatic system supervisory HAs are also included as design assumed HAs which are also evaluated in TA.) Operator's action also includes empirical and know ledged based HAs, from OER, and operation expert source, which may include knowledge-based tasks and cognitive activities required. The other consideration of operator's action includes procedural based HAs, administrative tasks which are necessary to execute in task sequence, including human interactions within and outside of MCR.

iii) Situation factors impose examinee to conduct operations under difficult situation by environmental condition, high-stress/workload (e.g., parallel tasks/events, unexpected/unrelated alarms, etc.)

### 2. Scenario built

Scenario built is a coordination scheme to organize each piece of conditions and HAs in a range of sampling dimension and integrate them to establish scenarios. It is important to ensure selected scenarios address all conditions and HAs which are defined in a range of sampling dimension. Operating procedure is helpful to verify this coordination since operators execute their tasks in accordance with procedures. It is also important to consider those scenarios should be practical and possible in reality.

After identifying scenarios, refining scenario script, i.e., initial condition, event sequence, staffing (examinee as well as test supporting staff), measurement tools etc. is defined.

## 3. Design Verifications

In accordance with NUREG-0711, rev.3, the V&V process is split to two parts of Design Verifications (DVs) and Integrated System Validation (ISV).

The two types of Design Verification: HSI Task Support Verification and HFE Design Verification. The HSI Task Support Verification is an evaluation to verify that the HSI provides the needed alarms, information, controls, and task support for personnel to perform their tasks, as defined by task analyses. The HFE Design Verification is an evaluation to verify that the design of the HSIs complies with HFE guidelines. The DVs can be conducted by desk-top base because they do not intend to evaluate human performance. Several elements of DVs can also be identified through the ISV as cause of performance degradation. However, ISV focus on human performance evaluation through the integrated system, dedicated DVs are necessary for overall evaluation of whether HSI inventory fulfills operator's needs to perform their tasks as well as the HSI design meets their style guide. The corrections of inconsistencies as defined by HEDs before ISV are necessary for efficient implementation of ISV.

## 4. Integrated System Validation

ISV is an evaluation, using performance-based test, to determine *an integrated system's design (i.e., hardware, software, and personnel elements) meets performance requirements and supports the plant's safe operation (NUREG-0711, rev.3).*

To make ISV successful and accomplish the goal, it is crucial to define performance measurement type (what aspects/characteristics should be measured), method (how to measure those measurement types), and criteria (provide pass/fail judgment basis) (IEEE Std. 845-1999). Performance measurement type breaks down performance measurement by interesting categories, which should be focused; primary tasks/secondary tasks, key factor to measure personnel tasks (i.e., time, accuracy, frequency, etc.,), situation awareness, workload, and anthropometric & physiological measures. Measurement method should be reliable, validate, repeatable, unobtrusive and objective. Various methods can be used to measure specific factor. For example, observation checklist, operation logs corrected by simulator, video camera monitoring operator's behavior can be leveraged for evaluating primary and secondary task accomplishment within the operation time required. To increase reliability, objectivity, and variability, multiple experts (maybe three) observation generates reliable, variable and objective data.

When applying the typical observation technique (i.e. walk-through/talk-through) in touch operation HSI system, additional observation tool may be applied to monitor examinee's behavior without disturbance and interruption by observer. Examinee's debriefing and questionnaire may be subjective but provide insight to identify the integrated system performance issues. Those questionnaires can be translated to quantitative data by formulating Likert scaling and statistic analysis. Performance measurement criteria should contain multiple bases. Comparison technique with benchmark/norm and task requirement (from TA under design assumption/performance constraint) provides objective (pass/fail) criteria. However, for overall judgment against acceptance of task performance evaluation (basically complicated and many factors involved), expert judgment and experimental practice may provide insight of cause of gap which can support to identify whether the gap is a real issue and should be correct. It is helpful to involve at least one operational expert who can interpret and represent examinee's behavior during ISV.

In case that plant function and scenario sequence can be comparable with the predecessor's plant system which is under operations and providing good performance data, the equivalent or improvement data in comparison with those in predecessor's plant system provide strong confidence for safety operation. In the case of a new plant application which does not have comparable data, if pass record with acceptable HSI design to resolve HEDs shows that the integration system does not provide any significant performance issues under possible situations, it would be a good assumption for safety operations. Subsequent process (human performance monitoring) will make sure to improve the integrated system performance during the life of plant.

## 5. Test design

Test design is a final step to integrate all logistics in the testbed and process. Scenario sequence identifies specific sequence and timeline of events, task measurement accommodations (observations, debriefing, etc.,) and the other logistics. Test procedure realizes test sequence to govern all test process in detail. Training for test examinee is necessary for their HSI system familiarization purpose. A pilot test may be conducted to make sure the scenario sequence can be conducted in accordance with the test procedure.

## Discussion of ISV Scenario Scope with Operating Process Element

Yun Goo KIM
*Central Research Institute, Korea Hydro and Nuclear Power Co., ltd, (ygkim.stpn@khnp.co.kr)*

The objects of this white paper are to a) provide general information of ISV in Korea and b) provide discussion issue about the scope of ISV scenario.

### Status of digital MCR in Korea

In Korea, 23 nuclear power plants (NPPs) are commercially operating and 5 NPPs (one OPR1000 and four APR1400) are under construction. Table 1 shows the status of APR1400 construction and planning. APR1400 is advanced pressurized water reactor which has compact workstation style control room. Shin Kori 3&4 is the first APR1400 plant and several types of HFE V&V for APR1400 have been completed such as preliminary verification, bottom up suitability verification, top down suitability verification, ISV, and final plant verification.

**Table 1 Status of APR1400 Construction and Planning**

| Plant name | Status | Commercial Operation | ISV |
|---|---|---|---|
| Shin Kori 3&4 | Under construction | 2015 | 2012.10 |
| Shin Hanul 1&2 | Under construction | 2017, 2018 | 2014.12 |
| Shin Kori 5&6 | Planning | 2019, 2020 | 2018 |
| Shin Hanul 3&4 | Planning | 2022 | - |

### ISV Practice in APR1400

Table 2 shows the general information of the ISV for Shin Hanul 1&2. The schedule of ISV was 3 weeks and 3 operator crew teams were participated to the ISV. Each operator crew team had one week test with 7 scenarios. Each scenario has 1~2 hours of operating length.

ISV scenario should include various plant operational conditions such as normal operation, I&C and HSI failures, transient and accidents. Specially, safety related operator actions such as emergency operating procedure execution and important human actions should be included in ISV scenario. However, there are practical limitations in ISV to include all safety related operator action. In Shin Hanul 1&2 ISV scenario, operational condition is screened out for integration to scenario. Even though it is screened out, one scenario still has various operational conditions.

**Table 2 ISV information of Shin Hanul 1&2**

| | |
|---|---|
| ISV schedule | 3 Weeks<br>2014.12.08 ~ 2014.12.27 |
| Test bed | Full scope Simulator |
| Participants | 3 Operating Crew Team |
| Number of Scenario | 7 scenario |

### Discussion about the scope of ISV and TSV scenario

Task support verification (TSV) has less limitation than ISV, because TSV can be performed in office with subject matter expert. Operational condition which is not included in ISV scenario can be verified by TSV with TSV scenario. However, TSV cannot validate integrated control room and there are difference between TSV and ISV. So, additional operational walkthrough is considered to demonstrate that the ISV with representative subset of operational condition ensures the validation of other operational condition which is not included in ISV. Table 2 shows the characteristics of TSV, ISV and operational walkthrough according to operation process element.

**Table 3 Operation process element and validation**

| Operation process element | Task Support Verification | Integrated System Validation | Operational walkthrough |
|---|---|---|---|
| Monitoring and detection | Verify that HSI provide the needed information(display and alarm) to operator | Validate that operator receive the provided information | Validate that the use of representative HSI can evaluate the use of other HSI |
| Situation Awareness | There is no direct verification. | Validate whether the operator aware the situation | Validate that the operator aware the situation with provided information, procedure, and training by using representative evaluation result |
| Response Planning | There is no direct verification. | Validate that the operator have response planning from procedure or knowledge | Operational condition with procedure<br>- Validate that procedure provide response activity<br>Operational condition without procedure<br>- Walkthrough with operation expert |
| Response Implementation | Verify that HSI provide the needed control means to operator. | Validate that operator can control the plant with provided control | Validate that the use of representative HSI can evaluate the use of other HSI |
| Team work (eg. communication) | There is no direct verification. | Validate that there are expected communication and team work | Walkthrough with operation expert |
| Result of operation (Plant status) | There is no direct verification. | Validate that the primary operation goal is met | Walkthrough with simulator |

Some of operation process element can be validate with representative ISV and walkthrough such as monitoring, detection, and response implementation. However, validity of operational walkthrough for other operational process element needs to be discussed. Following are questions for the discussion.

1. Should ISV scenario include all operator actions in normal operation, emergency operation, beyond design basis operation?
2. Is operational walkthrough needed for operational condition which is not included in ISV scenario?
3. How can operational walkthrough validate operational condition which is not included in ISV scenario?
4. Can operational walkthrough be included in task support verification?

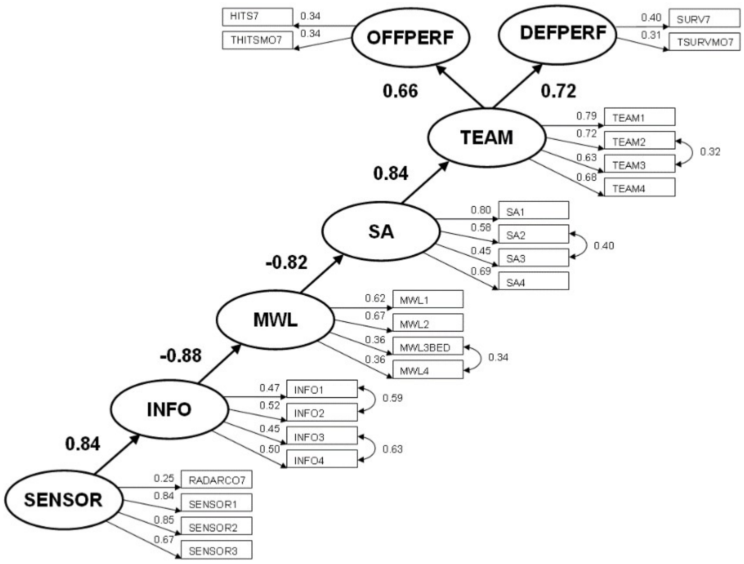
**Is Structural Equation Modeling as a component of the Integrated System Validation Toolbox an alternative to traditional statistical modeling approaches?** Swedish nuclear power plants have implemented many and several major plant modifications where the interaction of human-technology-organization (MTO) was essential. Two large control room changes have been implemented and one large control room change is underway. In Sweden is the Authority (SSM) not involved in validation but we have to trust in the licensee's validations. Therefore it is important for SSM that the validation is reliable and valid and analysis should be grounded on scientific methods, applicable for the evaluation of the MCR. It is also important that the method is presented transparent so it is possible to review.

SSM is in the startup phase of investigating an alternative statistical method for analyzing results from ISV, statistical modelling methods such as Structural Equation Modeling. The method and model is presented here, by our consultant from GEISTT AB.

**Structural Equation Modeling as a component of the Integrated System Validation Toolbox**
Data collection and data analysis is a constantly recurring challenge when you have the ambition of trying to describe and analyze a complex team work situation, for example as a part of an Integrated System Validation (ISV) process. Typically numerous and heterogeneous data sources are used in order to describe the work process and the interaction between human operators, technical systems and organizational aspects, which can create a dataset which is challenging to compile into useful information.

SSM is currently investigating the applicability of statistical modelling methods such as Structural Equation Modeling (SEM, e.g. Jöreskog & Sörbom, 1984[1]) as a component of the ISV analysis toolbox. SEM is a quantitative second generation statistical analysis method that combines the benefits of path analysis, factor analysis and multiple regression analysis. LISREL and AMOS are two of the more commonly known software packages that provide the computational capability of SEM. Experiences from the aviation domain, such as Castor (2009) will be used a starting point for SSM’s analysis. In Castor (2009[2]) the data from 308 simulated fighter aircraft engagements with four pilots and 24 variables measured during each engagement were analyzed and complied into the statistical model shown below. The database thus consisted of 1232 cases with 24 variables, generated by 37 pilots. The resulting model thus summarizes more than 700 hours of experience pilots’ complex behavior in an operationally valid environment. As a result of the statistical modeling effort, a structural equation model of how the chosen constructs relate to each other, and mediate effects between technical measures by a model of the operator, is proposed.

1. Jöreskog, K.G., & Sörbom, D. (1984). Analysis of linear structural relationships by maximum likelihood, instrumental variables, and least squares methods. Uppsala: University of Uppsala.
2. Castor, M. (2009). The use of structural equation modeling to describe the effect of operator functional state on air-to-air engagement outcomes. Linköping Studies in Science and Technology Dissertations 1251.
3. Linköping, Linköping university. http://liu.diva-portal.org/smash/get/diva2:207559/FULLTEXT01.pdf.



SENSOR = Sensor management
INFO = Usability of information
MWL = Mental Workload
SA = Situation Awareness,
TEAM = Teamwork quality
OFFPERF = Offensive Performance
DEFPERF = Defensive Performance

SEM is based on correlational statistics, i.e. the linear relationships between variables, and the common variance between the variables forms the basis for the analyses. SEM analyses and presents the degree of relationship between variables in terms of explained variance. A hypothesized model is tested statistically in a simultaneous analysis of the entire system of variables, to determine the extent to which the covariance or correlation matrix stipulated by the model, is consistent with the matrix based on the empirical data. If the statistical goodness of fit between the two compared matrices is adequate, the model is a plausible representation of the relations between variables that the model developer has specified.

Advantages

- SEM can combine different data types. Resulting models simultaneously analyze many statistical relations and present them in a comprehensive format
- SEM provides a goodness of fit value of the model vs the variance in the dataset, which can be used when comparing alternative models.

Disadvantages

- As all statistical methods it puts a number of requirements on the dataset.
- Requires some statistical understanding, sometimes beyond laymen's understanding.

Conclusion

- SEM is not a "silver bullet", but represents a powerful statistical analysis method which is useful for analysis of large and heterogeneous datasets from operational settings and processes.

Further questions

- martin.castor@geistt.com, +46 727 41 20 33
- yvonne.johansson@ssm.se, + 46 8 799 43 62

Workshop on Human Factors Validation of Nuclear Power Plant Control Room Designs and Modifications

# Recommended Research on Data Analysis and Drawing Conclusions

Nathan Lau
Grado Department of Industrial & Systems Engineering Virginia Tech
250 Durham Hall (0118)
1145 Perry Street
Blacksburg, VA 24061

Publish Date: 2015-02-03 **DRAFT**

## 1. Introduction

Integrated System Validation (ISV) is a critical activity to determine the adequacy of integrated control room design (i.e., hardware, software and personnel) in supporting safe and productive operations of nuclear power plants (NPPs) underdoing modernization and construction. Given the complexity of NPPs and extensive resource requirements for representative (or ecologically valid) performance testing, ISV presents substantial technical and practical challenges that can have serious safety implications. The Nuclear Energy Agency (NEA) has convened a group of experts to conduct this *Workshop on Human Factors Validation of Nuclear Power Plant Control Room Designs and Modification*. In this workshop, the experts examine the challenges and recommend research so that current methods could be improved and new methods could be identified to provide practically and technically sound means for achieving reasonable confidence in ISV test results.

The key organizers and contributors structure the workshop into two parts (i) challenges and (ii) recommended research to discuss four topical areas:

1. Defining the scope and objectives of control room validation
2. Rationale for selecting measures and acceptance criteria challenge
3. Test scenarios and test design
4. Analyzing the results and drawing conclusions

This *draft* white paper presents a perspective on *recommended research that supports analyzing test data and drawing performance conclusions* in the context of ISV of NPP control room design. The first section of the paper briefly describes the practical and technical constraints for representative performance-based testing in nuclear process control that shape the challenges in analyzing data and drawing conclusions for ISV. (Details of these challenges on data analysis and drawing conclusions are presented by other white papers.) The second section revisits the concept of conclusion validity for ISV that might help simplify the discussion on the approach to analyzing data and drawing conclusions. The penultimate section proposes two research areas – single-case experimental design and rater variability – that could have practical and technical merits for drawing conclusions on integrated control room design performance.

## 2. Analyzing Test Data and Drawing Performance Conclusions

ISV has many practical constraints in testing integrated operations of the control room that lead to many technical challenges in analyzing test data and drawing performance conclusions. These practical and technical constraints subsequently prompt many questions on the current ISV methodology.

### 2.1. Practical and technical constraints

The key, well-known practical constraint of performance testing is the labor requirement on two types of personnel – operator crews and experimental staff. For integrated testing of control room design, operator crews represent a key element but they are always in demand for other activities. Consequently, the number of crews and time of each crew allocated for ISV are limited, or at least kept to the minimum. The availability of operator crews is unlikely going to improve, posing continual challenges in generating performance data for drawing performance conclusions.

*Sample size.* The availability of operator crews for ISV poses three kinds of technical challenges for analyzing test data and drawing performance conclusions. The first two challenges concern conclusion validity due to inadequate sampling (for both qualitative and quantitative methods). Limited sample/number of crews participating in ISV activities may inadequately account for intra-crew variation (analogical to individual differences). Further, from a statistical perspective, crews should be randomly selected for appropriate generalization (i.e. random effect models) but this criterion is rarely satisfied strictly. Subjective and “objective” (i.e. statistical) correction could be applied to qualitative and quantitative data to moderate impact. However, the validity of correction methods is never examined for performance based testing in nuclear process control. In brief, inadequate sampling of operator crew could impact validity/confidence in drawing performance conclusions on integrated control room design.

*Scenario sampling.* The second challenge concerning conclusion validity is inadequate sampling of scenarios (as detailed in other white papers). Operator crews need to participate in a significant range of scenarios in order to provide the range of data to draw performance estimates or conclusion for “all operating conditions”. Further, from a statistical perspective, scenario selection, like operator crew recruitment, should be random for generalization but this criterion can be difficult to satisfy strictly. The limited availability of operator crews for performance testing implies that the number of sampled scenarios is also limited. Consequently, the generalization of test performance to operating performance has limitations, irrespective of quantitative and qualitative evaluation methods when operator availability is limited.

*Statistical power.* The final challenge associated with limited operating crew availability and thus performance data is the lack of statistical power for quantitative methods. Full-scope simulator studies typically require over 10 data points per experimental condition to indicate significant difference of medium-size effects. Additional data points are typically required for equivalence testing (that may be used in benchmarking studies). In brief, limited availability of operators constrains the traditional applications of (conventional) statistics for drawing performance conclusions.

The experimental (or performance testing) staff represents another set of practical constraints. A typical experimental staff involves human factors experts, simulation engineers and process experts. In particular, ample interaction time between human factors professionals and process experts i.e. experienced operators) are critical to develop effective scenarios and performance criteria. Further, performance measurements often involve some expert ratings. However, the availability of this interaction appears to be typically constrained or underestimated as process experts often have other duties and limited exposures to running performance testing from an ISV and experimental perspective. Further, human factors professionals do not have frequent opportunities to full-scope simulator evaluation while most experts are experienced with testing from a training and examination perspective. In brief, the interaction time between human factors professionals and process experts can have a major impact on quality of testing scenarios and measurements.

The dependence on the interactions between human factors professionals and process experts for quality ISV performance measurements has two technical implications. First, interactions between human factors professionals and process experts can drastically improve sensitivity and reliability of the performance measurements that provide the necessary statistical power for drawing conclusions. Second, performance data are often embedded with judgments or pre-conclusions of the process experts that may be unknown to the analysts. In addition, quality of expert judgments is a function of multiple factors, such as types of scenario event (e.g., within-design vs beyond design) and personal preferences (e.g., risk aversions). Consequently, measurement errors are not necessary constant, let alone individual differences/reliability between experts. Thus, the “confidence” of performance conclusions is actually a multivariate construct.

In nuclear process control, ISV also has two inter-related technical constraints that pose challenges in analyzing data and drawing conclusions. The first technical constraint is the combinatorial explosion of scenarios (or even scenario types) due to plant complexity. Formal methods do not exist to determine how interaction of components leads to qualitatively different scenario types and what portion of all possible scenarios does a particular set of scenarios cover. Thus, at least formally, the content validity or comprehensiveness of performance testing is difficult to assess. The second technical constraint is simulator fidelity of process behaviors for severe accident events such as Fukushima. The unknowns in limit the feasibility of direct adoption of well-established analysis methods in science.

## 2.2. Common concerns in analyzing data and drawing conclusions

The majority of data collection and analysis methods in science are developed (initially) with considerations of neither the industrial purpose (and safety implications) of ISV nor inherent constraints of the nuclear domain. In general, scientific studies produce knowledge or generate discussion for further testing and validation. Thus, science can accommodate “half-answers” to research questions. The nuclear industry clearly cannot accommodate “half-outcomes” to ISV projects. Further, methodological limitations often become impetus for further research. For instance, qualitative methods can focus on single participants to explore details and contexts with limited emphasis on generalization. Quantitative methods can focus on strict statistical or other criteria with large samples for validation. Methodological development in science often do not account for the constraints in the ISV of control room design.

The limited attention to the practical and technical constraints in many well-established methods to analyzing data and drawing conclusions ultimately prompts a series of research questions to support ISV in practice. The following questions summarize the ones posed by the workshop organizers:

- Are inferential statistics meaningful in the context of MCR validation? If so, what practical methods might be used to gather sufficient data during the human factors validation of the MCR to support the legitimate use of inferential statistics? If not, why not? What substitutes might be proposed as alternatives to traditional statistical modeling approaches? Are quantitative measures a requirement? If so, to what extent? Are inferential statistics a requirement?
- How might qualitative measures be analyzed or assessed to most effectively support achieving reasonable confidence?
- Is it possible (and desirable) to establish standardized acceptance criteria for the human factors validation of MCRs?
- Is there a rational basis for treating certain measures or data as only diagnostic while using others for pass/fail decisions during the human factors validation of the MCR?
- It is recognized that the validation process may result in changes to the integrated system (e.g. to the procedures, operator training, or a system interface). What should be the process or rationale for determining whether such changes should be evaluated through additional validation exercises?
- What should be the appropriate scope, format, and level of detail to be included in the final validation report?
- How should the validation results be aggregated and analyzed to determine the final validation conclusions?
    - Are inferential statistics meaningful in the context of MCR validation? If not, why not? What substitutes might be proposed as alternatives to traditional statistical modeling approaches?

The questions put forth belong to applied research, seeking ways to improve confidence in the data analysis and conclusion in the performance assessment of integrated control room operations given the practical and technical constraints in the nuclear domain (see Section 2.1).

## 3. A (Preliminary) Test Validity Perspective for ISV

The objective of improving confidence in data analysis and conclusions in performance testing of integrated control room operations connect closely to the longstanding research on test validity and validity generalization. In essence, this whitepaper is speaking to: What research can the nuclear community feasibly conduct to improve making valid ISV decisions from testing integrated control room operational performance? From this perspective, validity research can provide insights in the approach to and simplify the discussion on improving confidence in our ISV data analysis and conclusions.

Research questions on test and assessment validity has evolved over time (Murphy 2009). Prior to 1970s research focused on "Which forms of validity?" In the recent past between 1980 and 2000, research resolved the issue of "Is it valid" with meta-analyses. Presently, research is looking into "Validity for what?" from a multivariate perspective. This evolution, particularly the present phase, put ISV work into perspective. Prior to formulating any research programs to improve "confidence", the nuclear community must answer "Drawing conclusion for what?". The general answer to the question should be simple, though detailed versions may be contentious[1]. This white paper preliminarily proposes the following statement for drawing conclusions in ISV performance testing:

> "Conclusions from performance testing of integrated control room operations should help decide whether the integrated control room design can support reasonably safe plant operations over the requested licensing period (e.g., 20 years extension)"

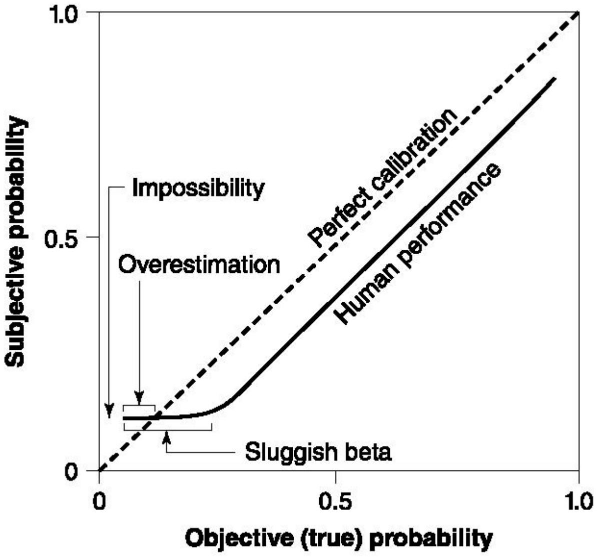
While the proposed answer is no epiphany to the workshop participants, the statement focuses data analysis and conclusions in terms of validity/confidence for **predicting safety over a number of years**.

This focus can help simplify or reframe the questions and discussions. Assuming the above statement is agreeable, ISV concerns with "consequential validity" and prediction. Thus, broadly speaking, all qualitative or quantitative data are collected and analyzed to make (or to become confident in making) **inferences on future performance**. Though the traditional criteria may be impractical, many principles in "inferential statistics" (e.g., type I and II errors) remain essential for establishing confidence in ISV performance conclusions. From this perspective, all evidence that could support the prediction should be admissible and documentation needs to reflect consequential validity of the data, analysis and results.

However, individual pieces of evidence are not equal in merits and confidence. Besides varying importance for different performance categories (discussed in other white papers), both qualitative and quantitative data likely contain varying levels of measurement errors that affect data analysis and conclusions. Other than questionnaires administered to the participants, process experts are involved in data collection and/or analysis methods. Measurement errors for a specific measure within a single process expert are probably not constant. For instance, process experts might be more accurate and reliable in estimating/judging operator performance for within-design basis than beyond-design-basis scenarios. The rationale of this statement can be induced from gross over- and under- estimation of event probabilities for extremely common and rare events (Wickens et al. 2012). Further, most people consistently underestimate probabilities for frequencies in the mid-range (see Figure 1). Discovering the tendencies of process experts in performance interpretation is also applicable to qualitative as well as quantitative data analysis. In summary, improving knowledge on measurement (and interpretation) errors associated with process experts for individual measures and conditions can improve the confidence in any performance conclusions.

[1] Test validity is a multivariate construct; thus, one central statement, no matter the level of emphasis, is insufficient to represent the full concept for ISV in the nuclear domain. One pertinent area missing in the statement is the resource requirements that must be balanced with confidence in the predictions. If resource requirement is completely ignored for perfect predictability, then all NPPs would not be licensed and such consequence invalidates the target outcome of performance testing.

**Figure 1: A hypothetical weighting function. Extracted from Figure 8.8 of Wickens et al. (2012)**



If the assumption that measurement errors vary with (some) scenario types is valid, the methods of data collection and analysis deserve strategic selection. For instance, small measurement errors for a particular type of scenarios afford few data points to achieve the necessary confidence (though not necessarily statistical significance). On the other hand, large measurement errors for some scenario types deserve increased sampling and caution. Further, operator performance for handling familiar events may orient towards quantitative methods while performance for handling unanticipated events may orient towards qualitative methods. (This idea should be subjected to debate.) Empirical research on correction factors pertinent to operator performance assessment in the nuclear domain is absent. Given the stellar safety records in the nuclear industry, “confidence” in ISV decisions (according to the author’s opinion) is likely stemming from conservative safety culture and deep domain knowledge of the nuclear industry. Safety culture and domain knowledge are likely to influence ISV measurements and decision validity in the future.

Answers to many questions put forth by the workshop organizers are not being mentioned, let alone discussed from the test validity perspective in this white paper. Nevertheless, this whitepaper highlights that (i) the conclusions of performance testing for ISV is about (at least in part) making inferences/predictions on safety, and (ii) the confidence in the predictions is highly dependent on the process expert(s) supporting the testing activities.

### Recommended Research

In the typical circumstances when data points are fewer than necessary to provide conclusive (conventional) inferential statistics, the industry are faced with the question on whether the collected quantitative and qualitative evidence provides sufficient confidence in predicting reasonably safe operations for the licensing period. The judgment on the confidence in the collected data would in part rest on our understanding of the measurement errors or variability in the data collection and analysis methods. That is, our knowledge in methods (and the domain) can be and probably is being applied to moderate the confidence levels in individual performance results (c.f., probability estimates in human reliability analysis). Thus, research that can provide an empirical foundation for estimating the “errors” or variability of various performance testing conditions and measures would improve confidence in ISV conclusions and decisions. The improved confidence stems from an empirical basis for moderating or correcting the confidence levels for individual performance findings (or claims).

A feasible research strategy to study measurement variability under different testing conditions would be invaluable for building an empirical basis that assesses merits of ISV performance results from small-sample size testing. The proposed strategy to acquire knowledge on measurement variability is to collect (or simply record) qualitative and/or qualitative performance data during simulator training required for licensed operators several weeks per year (e.g., five weeks in the US). In fact, this strategy is already proposed by a recent HRA research program called Scenario, Authoring, Characterization, and Debriefing Application (SACADA; James Chang et al. 2014)[2]. This strategy is feasible because "much of the work" for the performance testing is already being (and must be) done. The key missing element in simulation training sessions at nuclear power plants appear to be formalizing the measures and recording the data. If the utilities would implement well-established human performance data collection methods and share the data in their simulation training sessions, sufficient data can be feasibly collected to study human performance measurement errors, particularly in expert judgment variability across multitude of test factors (such as scenario types). Further, the "data bank" can provide a reference performance level that put a particular data point collected in ISV performance testing into context. That is, the collected data would have a corresponding performance and variance level for comparison.

Collecting and recording human performance data during simulator training sessions offer five benefits that could contribute to the confidence in drawing performance conclusions in ISV.

1. Collecting human performance data provides reference performance and variability levels for many common scenarios of performance tests, which individually cannot provide the sample size for generalization. This reference performance and variability levels offer an empirical basis for qualitatively and quantitatively evaluating the ISV results, thereby improving the confidence (or consequential validity) in the ISV decision. In addition, a priori hypothesis can be formulated with respect to the reference performance levels if necessary (though unlikely given resource constraints).
2. In relation to the first benefit, the available data provide indication on when expert judgment and corresponding interpretation become less reliable, yielding weak performance prediction. Isolating less reliable performance results and thus prediction can improve the ISV conclusion and decision validity.

   By identifying the conditions with poor reliability, research and ISV efforts can be allocated accordingly. For instance, ISV efforts may focus testing on beyond-design-basis once comparable performance and variability levels become apparent for within-design-basis scenarios. Alternatively, the ISV can mainly target new failure modes that does not exist in the original plants.

[2] This approach is similar to the SHRP2 naturalistic driving study that instrumented 3000 vehicles to observe and collect data on ordinary people about their driving behaviors. The data bank supports calculating odds ratio of a particular behavior leading to crashes (relatively rare events).

3. Standardization in collecting and recording human performance data in already formal simulator training sessions provides benchmark performance for plants undergoing modernization. The benchmarks enable other data collection and analysis methods, including single-case experimental designs (SCED) that rely on visual inspection of performance data (Rizvi and Ferraioli 2012). The medical field employs SCED frequently to study intervention effective but the method requires careful measurements of baseline behaviors. If utilities undergoing modernization are actively collecting performance data during simulator training sessions, SCED can be effectively employed for ISV. Some research on this method in the nuclear domain is likely warranted. In any case, collecting human performance data during simulator training can provide the additional data to facilitate test designs that can improve confidence in drawing performance conclusions.
4. Another benefit mainly relevant for modernization projects is the exposure of process experts to using or implementing current human performance data collection methods. As mentioned, the quality of human performance measurements depends heavily on working with process experts. Increased exposure to using and working with human performance measures reduces time demand to produce quality data when performing ISV. Quality data improve confidence in performance conclusions (and statistical power for quantitative analysis).

## 4. Conclusion

Given that limited data that can be feasibly collected from a specific performance test, performance conclusions in ISV often cannot practically be drawn according to criteria designed for scientific pursuit (mostly in academia). In the nuclear industry, ISV results for predicting future performance become heavily dependent on the domain experience of the process experts and other professionals to “fill data void”. Thus, one strategy to increasing the confidence in performance conclusions would be to learn about the test conditions in which experts make reliable or unreliable performance estimation and prediction. Further, reference performance levels would be invaluable to put individual performance data point into context. Studying measurement variability and providing reference performance levels can be **feasibly** accomplished by collecting qualitative or quantitative performance data during simulator training sessions necessary for licensing individual operators. That is, simulator training sessions can help generate a human performance data bank that could help interpret individual performance data points collected from ISV performance testing. This data bank would help estimates the level of confidence in a specific ISV finding and thus the specific performance prediction.

## References

James Chang, Y., Dennis Bley, Lawrence Criscione, Barry Kirwan, Ali Mosleh, Todd Madary, Rodney Nowell, Robert Richards, Emilie M. Roth, Scott Sieben, and Antonios Zoulis. 2014. "The SACADA database for human reliability and human performance." *Reliability Engineering & System Safety* 125 (0):117-133. doi: http://dx.doi.org/10.1016/j.ress.2013.07.014.

Murphy, Kevin R. 2009. "Validity, Validation and Values." *Academy of Management Annals* 3 (1):421-461. doi: 10.1080/19416520903047525.

Rizvi, Shireen L. , and Suzannah J. Ferraioli. 2012. "Single-case experimental designs." In *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological.*, edited by Harris Cooper, Paul M. Camic, Debra L. Long, A.

T. Panter, David Rindskopf and Kenneth J. Sher, 589-611. Washington, D.C., USA: American Psychological Association.

Wickens, Christopher D., Justin G. Hollands, Raja Parasuraman, and Simons Banbury. 2012. *Engineering psychology and human performance*. 4th ed. Upper Saddle River, NJ: Pearsons.

# Human Factors Validation of Nuclear Power Plant Control Room Designs and Modifications

Valerie J. Gawron PhD
The MITRE Corporation*

## Challenge Issue and Proposed Resolutions - Critical Considerations in Defining Validation Objectives and Their Impact on Confidence

Human factors is only one component of human systems integration (HSI) that is a critical consideration in validation of any complex system. As shown in Figure 1 and defined in Table 1, HSI covers all aspects of the human that affect human system performance each of which must be part of any system validation. The Air Force Human Systems Integration Handbook (Directorate of Human Performance Integration, 2014) contains excellent checklists and guidance for evaluating HSI issues throughout a system life cycle.

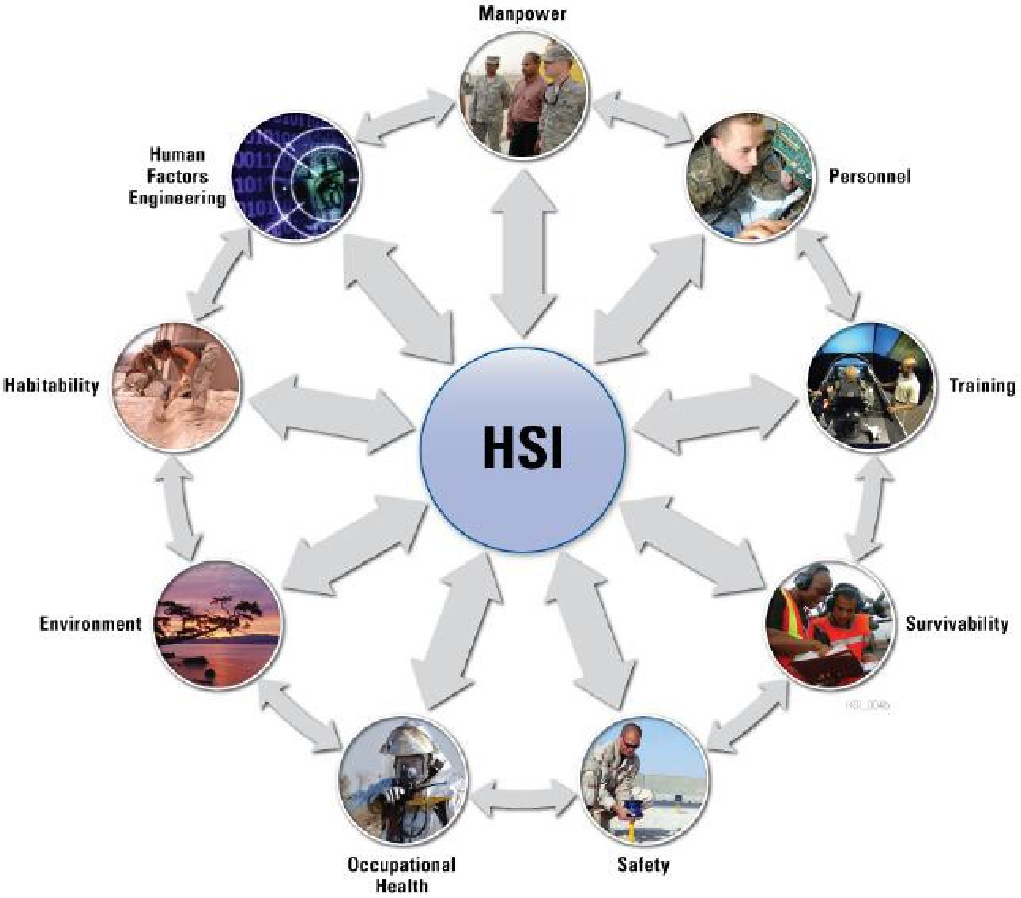


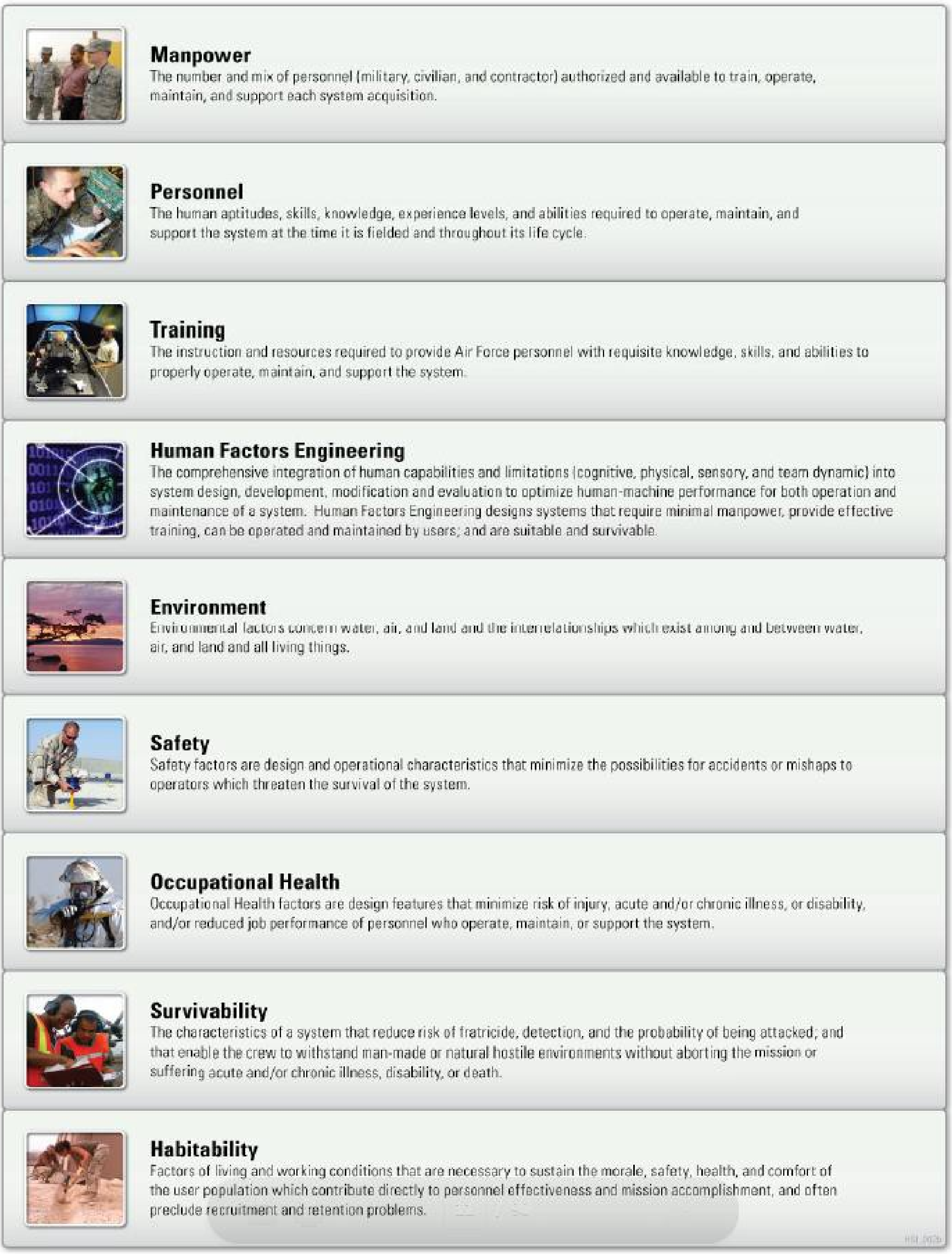
**Figure 1. HSI Integrated the Domains**
(Directorate of Human Performance Integration, page 40)

* Approved for public release; distribution unlimited. Case number 15-0246

© 2015 The MITRE Corporation. All RIGHTS RESERVED.

**Table 1. HSI Domains and Definitions**

(Directorate of Human Performance Integration, page 11)



**Manpower**
The number and mix of personnel (military, civilian, and contractor) authorized and available to train, operate, maintain, and support each system acquisition.

**Personnel**
The human aptitudes, skills, knowledge, experience levels, and abilities required to operate, maintain, and support the system at the time it is fielded and throughout its life cycle.

**Training**
The instruction and resources required to provide Air Force personnel with requisite knowledge, skills, and abilities to properly operate, maintain, and support the system.

**Human Factors Engineering**
The comprehensive integration of human capabilities and limitations (cognitive, physical, sensory, and team dynamic) into system design, development, modification and evaluation to optimize human-machine performance for both operation and maintenance of a system. Human Factors Engineering designs systems that require minimal manpower, provide effective training, can be operated and maintained by users; and are suitable and survivable.

**Environment**
Environmental factors concern water, air, and land and the interrelationships which exist among and between water, air, and land and all living things.

**Safety**
Safety factors are design and operational characteristics that minimize the possibilities for accidents or mishaps to operators which threaten the survival of the system.

**Occupational Health**
Occupational Health factors are design features that minimize risk of injury, acute and/or chronic illness, or disability, and/or reduced job performance of personnel who operate, maintain, or support the system.

**Survivability**
The characteristics of a system that reduce risk of fratricide, detection, and the probability of being attacked; and that enable the crew to withstand man-made or natural hostile environments without aborting the mission or suffering acute and/or chronic illness, disability, or death.

**Habitability**
Factors of living and working conditions that are necessary to sustain the morale, safety, health, and comfort of the user population which contribute directly to personnel effectiveness and mission accomplishment, and often preclude recruitment and retention problems.

**Manpower** – Whan and Long (1984) identified nuclear power manpower and training requirements and the interdependency of the two. The classic work is still applicable today. Manpower affects not only training requirements but also safety. In June 2012, Entergy postponed a key safety drill at the Pilgrim Nuclear power Plant due to lack of manpower.

**Personnel** – Some have long predicted that qualified manpower for the nuclear industry will be in short supply as many experienced personnel reach retirement age and the number of graduates declines. The Organization for Economic Cooperation and Development (OECD) Nuclear Energy Agency identified this issue two decades ago (Zarimpas, 1993) and the problem has not diminished (Business Wire, 2012). If there is a shortage of personnel to operate the plats who will be the testers?

**Training** – The range of training requirements is huge and covers "operation and maintenance, fuel processing, spent fuel management, and decommissioning" (Business Wire, 2012, 27). Validation of nuclear power plant control room designs and modifications must include this entire range of tasks, training types, and levels of expertise.

**Human Factors Engineering (HFE)** – The greatest challenge in validating complex systems is understanding the effects of automation on human performance. John Kenneth Galbraith (1967) stated "We are becoming the servants in through, as well as in action, of the machines we have created to serve us". One item that is not included in the Air Force HSI Handbook but is critical to validation of nuclear power plants is maintainability. This type of **maintainability** is described in a classic article by Seminara and Parsons (1982).

**Environment** – A classic article (Carach, 1980) describes the working environment in nuclear power plants while another (Connell, 1975) describes the effects of marine environment on nuclear power plant design.

**Safety -** Of critical import for nuclear power plants is safety. One variable that greatly impacts safety is the safety culture. Meshkati (1995) identified three conditions necessary for a safety culture in nuclear power plants: "1) a thorough understanding of the dimensions of the local national culture, 2) determination of the extent of their match with those of organizational culture of the installation, and 3) their compatibility with the prescribed requirements for safety culture" (page 286).

**Occupational Health** – Fukushima Dai-ichi demonstrated the occupational health challenges of a nuclear disaster (Wada, Yoshikawa, Hayashi, and Aizawa, 2012). But the other side of the coin the R.E. Ginna Nuclear Power Plant has been recognized for its excellent occupational safety and health record (Energy Weekly News, 2009).

**Survivability** – A continuing problem in validation of nuclear power plant control room designs and modifications is the management of ageing and legacy equipment (International Atomic Energy Agency, 2000). Further the range of systems to increase

the probability of survival during and after an incident at a nuclear power plant is immense from flame arresters (Hong, Song, Kim, and Chang, 2006) to communication system reliability (Makhutov, 2013) and terror proofing (Pfister, 2007).

**Habitability** – An excellent review of nuclear power plant control room habitability (Transactions of the American Nuclear Society, 2003) provided an expert panel's assessments of critical design issues.

**Additional Consideration** – Although it is affected by all aspects of HSI there are several unique aspects to automation that should be addressed during the validation of nuclear power plant control room designs and modifications. I reviewed the accident and incident data in aviation, factories, and health care and identified five areas for automation-induced error research: vigilance, skill retention, failure detection, automation-induced complacent, and situation awareness (SA). These are described in Appendix A drawn from a case study report completed in 2002.

**References**

Carach, J. (1980). Working environment in nuclear power plants. *Pracovni Lekarstvi, 32*(4), 147-153.

Connell, C.W. (1975). Effect of marine environment on nuclear power plant design. Elsevier, 625 – 628.

"Constellation Energy; Constellation Energy's R.E. Ginna Nuclear Power Plant Recognized for Safety Program by Occupational Safety and Health Administration." Energy Weekly News (2009): 121. ProQuest. Web. 27 Jan. 2015.

"Entergy Postpones Key Safety Drill at Pilgrim Nuclear Power Plant due to Lack of Manpower." PR NewswireJun 15 2012. ProQuest. Web. 27 Jan. 2015.

Galbraith, J.K. (1967) The New Industrial State. Princeton University Press.

Hong, Seong-Wan, Song, Jin-Ho, Kim, Hee-Dong, Chang, Soon-Heung. (2006). The applicability of a quenching mesh as a hydrogen flame arrester in nuclear power plants. *Nuclear Technology, 153*(1), 89-99.

International Atomic Energy Agency (2000). Management of Ageing of I&C Equipment in Nuclear Power Plants. IAEA Publication TECDOC-II47, Vienna, Austria.

Makhutov, N. A. (2013). A criterion base for assessment of strength, lifetime, reliability, survivability, and security of machines and man-machine systems. *Journal of Machinery Manufacture and Reliability, 42*(5), 364-373. doi:10.3103/S1052618813050075

Pfister, Bonnie. "NRC Seeks to Terror-Proof Nuclear Plants." McClatchy - Tribune Business News: 1. Jan 18 2007. ProQuest. Web. 27 Jan. 2015

"Research and Markets: 2012 Report - Work Force Development to be a Top Priority for the Growth of Nuclear Industry." Business Wire Apr 27 2012. ProQuest. Web. 27 Jan. 2015.

Seminara, J.L, Parsons, S.O. (1982). Nuclear power plant maintainability. Applied Ergonomics, 13(3), 177-189.

U.S. nuclear power plant control room habitability - panel. (2003). *Transactions of the American Nuclear Society, 88*, 59.

Wada K, Yoshikawa T, Hayashi T, Aizawa Y. Emergency response technical work at Fukushima Dai-ichi nuclear power plant: occupational health challenges posed by the nuclear disaster. Occupational And Environmental Medicine [serial online]. August 2012; 69(8):599-602. Available from: MEDLINE Complete, Ipswich, MA. Accessed January 27, 2015.

Whan, Glenn A, Long,Robert L. (1984). Nuclear power manpower and training requirements. Energy, 9(9-10), 919-928.

Zarimpas, Nicholas. "Qualified Manpower for the Nuclear Industry." Organisation for Economic Cooperation and Development. The OECD Observer.181 (1993): 8. ProQuest. Web. 27 Jan. 2015.

**Recommended Publications**

Directorate of Human Performance Integration Air Force Human Systems Integration Handbook Planning and Execution of Human Systems Integration. Brooks City- Base, RX: Human Performance Optimization Division (http://www.wpafb.af.mil/shared/media/document/AFD-090121-054.pdf)

Meshkati, N. (1995) Cultural context of the safety culture: A conceptual model and experimental study in safety culture in nuclear installations. Proceedings of the International Topical meeting on Safety Culture in Nuclear Installations, Vienna, Austria, 286 – 297.

### 1.1 The Watching - A Review of the Vigilance Research

During WWII an unusual performance phenomenon was noted among radar operators: specifically, the probability of detecting a signal decreased dramatically over time-on-watch. This decrement was greatest when:

1. The signal duration was short.
2. The probability of a signal was low.
3. The signal intensity was low.
4. The signal was simple rather than complex

A number of theories were developed to explain these findings. The most well know of these is the Theory of Signal Detection (TSD). TSD was introduced over forty years ago (Swets, Tanner, and Birdsall, 1961). Since that time it has been one of the most widely used theories of human behavior. There are two main reasons for the theory’s popularity. First, TSD is a model of human detection in noise and, therefore, applicable to many real-world situations. Second, TSD is a normative model and thus describes how the human should act. As such, TSD is useful in developing criteria of optimum performance against which actual performance can be compared. From this comparison, suggestions can be made for improving a system.

TSD is a rather simple and straightforward theory. It proposes that there are two states in the real world: noise alone or a signal in that noise. Each of these states gives rise to neural activity. The activity from each of these states is the same at certain intensities, and the probability that a given intensity is due to either state is hypothesized to describe two overlapping normal curves (see Figure 12). In the area of overlap, detection becomes not only a sensory problem but a decision-making one, as well (Swets, 1964; Swets, et al., 1961).

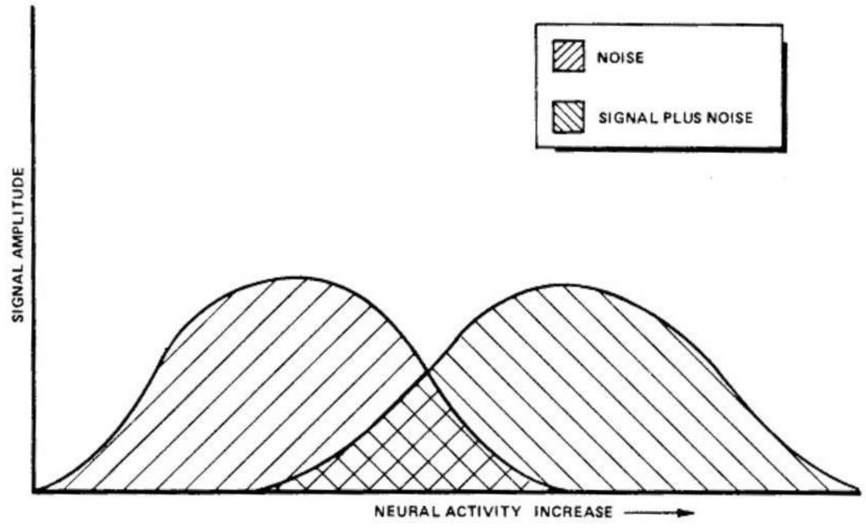


**Figure 12 Two States of the World**

The decision-making aspect of the detection process is described by the TSD concept of the likelihood ratio (λ). This is the ratio of the probability that a given intensity of neural activity resulted from the occurrence of a signal to the probability that the intensity resulted from noise alone. Above a certain value of λ, the operator will decide that the signal did occur and below this value he or she will decide that he neural activity is due to noise alone. The value of λ that serves as a yes-no decision demarcator is called β (see Figure 13).

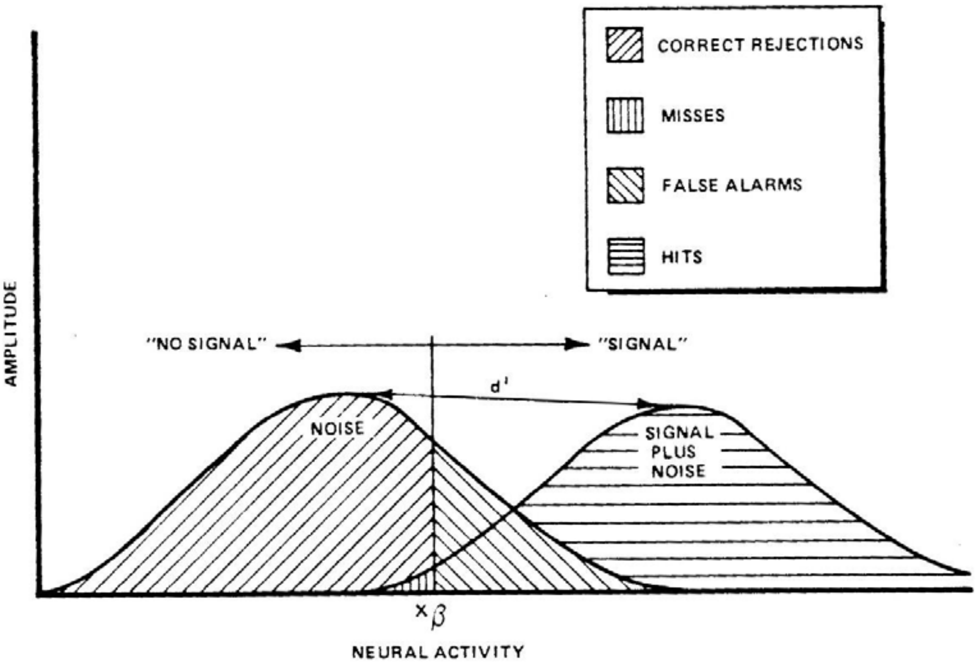


**Figure 13 λ as a Function of Neural Activity**

Given the decision scheme proposed by TSD, there are four possible outcomes (see Figure 13). First, an operator may decide that a signal occurred when one actually did – a hit. Second, he or she may decide that no signal occurred when in fact none did – a correct rejection. These two outcomes are clearly desirable. However, there are two other outcomes that are not desirable. An operator may decide that a signal occurred when one did not – a false alarm. Finally, he or she may decide that no signal occurred when in reality, one did – a miss.

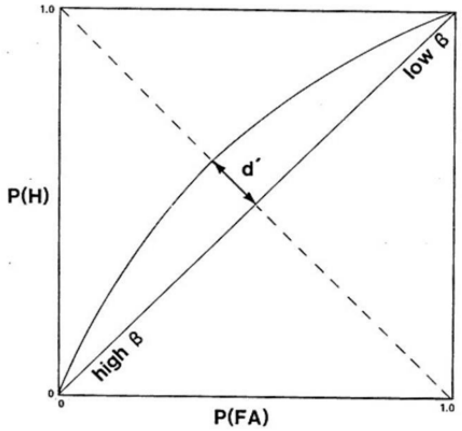
The probability of each of these outcomes is described by the areas of the noise and signal-plus-noise curves allotted to each. As Figure 13 clearly demonstrates, the frequency of occurrence of these outcomes is determined by β. A high or conservative β is associated with a large number of missed signals but a small number of false alarms. On the other hand, a low or risky β will result in a large number of hits but also a decrease in correct rejections.

In both cases there is a tradeoff between desirable and undesirable outcomes. It is at this point that the normative aspect of TSD becomes useful since it can describe an optimum value. This value is defined in terms of the probabilities of the occurrence of noise (P(N)) or signal (P(S)), the values of correct rejections (V(CR)) and hits (V(H)) and finally, the costs of false alarms (C(FA)) and misses (C(M)). The equation defining the optimum β value is presented below.

$$\text{Optimum } \beta = \frac{P(N)}{P(S)} \times \frac{V(CR) + C(FA)}{V(H) + C(M)}$$

The operator's actual performance can be compared against this optimum and suggestions made for the improvement of performance. This comparison is made by manipulating the operator's perception of values and costs of the decision outcomes and plotting the resulting probabilities of hits (P(H)) versus false alarms (P(FA)). This function is called a receiver operating characteristic (ROC) curve. The curve is especially useful since the distance of the curve to the positive diagonal of the graph yields a measure of the operator's sensitivity irrespective of decisional manipulations. This measure is termed d' (see Figure 14).

**Figure 14 ROC Curve**



d' also corresponds to the distance between the noise and signal-plus-noise curves (see Figure 14). As such, it is a measure of the difficulty of the task. As d' increases, the amount of overlap between the two curves decreases and the detection task becomes easier. Similarly, as d' decreases, the amount of overlap increases, making the detection task more difficult. Goodenough and Metz (1974) suggested that vigilance decrement is in reality a decrease in d'.

The theory is not without its critics, however. Broadbent (1971) questioned the verity of four TSD assumptions. First, TSD proposes that the distributions of the likelihood ratios are normal and have equal variance. This may not be true. Second, TSD assumes that an observer can arrange the perceived sensory evidence along a single dimension. Given the diversity of sensory inputs and the human's inability to aggregate evidence over time, this assumption is also questionable. Third, it is assumed that the observer knows the λ value for every possible experienced event. However, the human is an imperfect aggregator of evidence. Fourth, TSD assumes that the observer is always rational. This is again questionable. TSD, however, has served as a good model for research.

Another theory, proposed by Welford, stated that the vigilance decrement was the result of a decrease in attention (arousal) over time. Four independent lines of research supported this theory. First, psychophysiological recordings of operators during vigilance tasks indicated that poorer detection performance was associated with lower physiological arousal (Blakeslee, 1979; O'Hanlon, 1979). Second, stimulants such as nicotine and caffeine that raise an operator's physiological arousal decreased the magnitude of the vigilance decrement. Third, loud auditory stimulation, which also results in increased arousal, decreased the magnitude of the vigilance decrement. Finally, vigilance performance under the threat of electric shock was superior to that without the threat of shock (Kennedy and Coulter, 1975).

Welford's theory, however, could not explain the vigilance decrement that occurred when newly trained operators were on watch. The decrement they exhibited was much greater than that shown by experienced operators, and yet the novices seemed to be the more highly motivated and aroused of the two groups. To explain this phenomenon, Baker proposed another theory. It was based on the operator's expectancy of the signal occurrence. During training, operators are typically exposed to many signals in a short period of time. This high signal rate is meant to efficiently enhance the trainee's ability to recognize signals. The high signal rate, however:

1. Causes the trainee to adopt a conservative decision rule, i.e., a stimulus must be strong to be considered a true signal; and
2. Creates the expectancy of a high signal rate in the real world.

The operator thus misses weak signals especially as time-on-watch progresses. Further, the newly trained operator confronted with a low probability task, gradually becomes more familiar with the signal probability in the real world and attempts to match his or her response probability with it (Craig, 1978). It was found that the vigilance decrement of newly trained operators could be reduced if the probability of a signal was minimized during training.

The final vigilance theory to be described is the Vicious Circle Hypothesis. This hypothesis suggests tha an observer's decisional rule (β) is determined by a subjective estimate of the probability of a signal (P(S)). But this is based on the probability of the observer saying yes and this in turn is based on the observer's β. Over time the observer will become more conservative. This causes him or her to miss weak signals and to decrease his or her subjective estimate of signal probability. This decreases the probability of saying yes, which makes the observer become even more conservative, hence the vicious circle.

The research to date has not clearly differentiated the strengths of the four theories presented above. This is partially because the research, especially that sponsored by the military, has been directed at more applied problems. The applied research suggests that vigilance decrement can be decreased by:

1. Providing the operator with knowledge of results, i.e., the number of signals missed and the number of false alarms (Poulton, 1973),
2. Increasing the operator's motivation,
3. Providing the operator with a more accurate (i.e., real-world) estimate of signal probability (Craig, 1980),
4. Periodically exposing the operator to sample signals to maintain an accurate image of the signal characteristics in his or her memory,

5. Introducing false signals, and
6. Training the observers to use either fixed or completely random search patterns rather than alternating patterns (Fisk and Schneider, 1980).

The first four methods are difficult to implement in the real world. The fifth is akin to crying wolf and can give the operator an inaccurate estimate of system status. But, Mitta and Folds reported empirical date that high false alarm and hit rates as well as short detection latency enhance operator performance. Their data were collected in an automated traffic incident detection system. But vigilance performance even in high signal-rate environments (e.g. 30 signals/minute, Parasuraman, 1980) can be degraded over time. Also, the detection latency to real signals is greatly increased by the presence of non-relevant signals (Thackray, Bailey, and Touchstone, 1979). Finally, introducing false signals was originally meant to increase operator stimulation thus avoiding under load. However, Warm, Dember, and Hancock (1996) cite a number of experiments to support their conclusion that “under load is a myth: rather than being under stimulating, vigilance tasks are exacting, capacity-draining assignments that are associated with a considerable degree of mental demand and frustration” (p. 195).

Finally, a sobering finding is that vigilance of pairs can also degrade. For example, Wiener (1980) identified vigilance problems of both the pilot and the ATC to be at fault in midair collisions. There is clearly yet much to be learned about human vigilance and automation.

## 1.2 The Doing –A Review of the Skill Retention Research

The problem of how to maintain skills in the most cost effective manner has been plaguing system developers for decades. This problem is especially prevalent in the military where some tasks may not performed for real in anyone’s career but the soldier must remain prepared. It is also prevalent in automated systems in which the manual skills may rarely if ever be needed. In an early study, Van Dusen and Schlosberg (1948) examined the effect of retention interval (1, 7, or 28 days) on performance of perceptual motor tasks. There were significant decreases in performance after 7 and 28 days. Also in an early study, Ammons (1951) reported that massed practice resulted in superior performance in a rotary pursuit-tracking task on the following measures: mean duration of hits, mean duration of misses, and number of hits. However, distributed practice resulted in longer total stylus-target contact time. However, Catalano (1978) reported that a one-minute rest resulted in both greater time on target and absolute error in a tracking task. The author attributed the latter to a warm-up decrement.

Neuman and Ammons (1957) studied the effects of retention interval (1 minute, 20 minutes, 2 days, 7 weeks, and 1 year) on the number correct in the perceptual motor task. They concluded that initial retention test performance was progressively worse as the retention interval increased. Further, the amount of retraining required increased as the length of the retention interval increased. In a follow-on study, Ammons, Farr, Black, Neumann, Day, Marion, and Ammons (1958) evaluated the effect of degree of training (5 versus 30 trials) and length of retention interval (1 minute, 1 day, 1 month, 6 months, 1 year, or 2 years) on a perceptual motor task. The amount of training resulted in decreased task completion time. As the length of the retention interval increased, the task completion time also increased. The authors concluded that proportionately fewer trials are required by subjects having greater training to achieve former performance levels. Their subjects were 538 male undergraduate students. In a second experiment with 465 male undergraduate students, the authors looked at the effect of amount of training (1 versus 8 hours) and duration of the no-practice interval (1 day, 1 month, 6 months, 1 year, or 2 years) on time on target in a compensatory two-dimensional tracking task. As expected, increased practice resulted in increased time on target. However, after two years there was no difference between 1- and 8-hour practice groups. The amount of skill retention decreased as time increased. However, in a similar study, Adams and Reynolds (1954) examined the effect of massed versus distributed practice on performance of a rotary pursuit-tracking task. Their subjects were Air Force recruits. They reported no difference in type of practice on performance.

Fleischman and Parker (1962) compared performance on a complex, continuous tracking task over 1, 5, 9, 14, and 24 months. They reported retention of the task was “extremely great”. However, there was a decrement in performance after the 14th month. They concluded that skill retention was determined by initial proficiency rather than by the type of training (formal versus informal). Further, there was no significant difference in performance after 1 week between subjects who had massed versus distributed practice.

Brown, Briggs, and Naylor (1963) reported a study investigating different types of rehearsal on reaction time and three-dimensional tracking tasks. They concluded that the amount of original training, if sufficiently long, can ellipse any positive effects of rehearsal. Their subjects were 126 male undergraduate students. In a follow-on study, Buckout, Naylor, and Briggs (1963) used the same two tasks to study the effects of length of training (one versus three weeks), visual noise (present or absent), and feedback sensitivity (95 decibels (dB) tone triggered by either large or small errors). The subjects (142 undergraduate students) with more training had better performance after the retention interval than subjects with less training. Visual noise resulted in significantly greater tracking error. However, subjects trained with visual noise had fewer tracking errors and better reaction time performance after the retention interval than subjects who were trained with no visual noise. There was no effect of type of feedback. In a similar study, Melnick (1971) compared the performance of 80 male undergraduate students who had received 0, 50, 100, or 200% over learning practice. The retention intervals were one versus four weeks. The task was the stabilometer. Melnick (1971) reported that immediate recall of the task was facilitated by over learning. After 4 weeks, subjects who received 200% over learning had better retention that subjects who had 0% over learning.

Using a similar task, Thompson, Wenger, and Bartling (1978) measured word recall from a list in a series of three experiments. Multiple presentations helped more with longer (48 hours) than shorter (20 minutes) retention intervals. These results are based on three experiments with undergraduate students.

Naylor and Briggs (1963) also examined the effects of type of rehearsal: whole task, temporal, spatial, or no rehearsal. The subjects were 68 undergraduate students. The task was reaction time. Performance was more accurate with rehearsal. Whole-task rehearsal resulted in the best performance. In a series of two similar experiments, Naylor, Briggs, Brown, and Reed (1963) examined the same types of rehearsal but on a procedural task. The effectiveness of the rehearsal was not in the expected order but rather the most effective to least effective rehearsal was as follows: part task, simplified, whole task, and no rehearsal. In the second experiment, for a tracking task, whole-task rehearsal resulted in superior performance to part-task or no rehearsal. The effect of rehearsal decreased with an increase in the amount of original training. Whole-task rehearsal was superior up to five days of training; part-task after eight days of training; but after 10 days of training there was no rehearsal effect. Naylor, Briggs, and Reed (1962) also looked at the amount of training. Subjects were required to perform a procedural task with either a high or low degree of task organization. The authors concluded that task organization has greater influence on performance with less amount of training. Naylor, Briggs, and Reed (1968) examined the effects of training time (2 versus 3 weeks), retention interval (1 versus 4 weeks), and task coherence (sequential versus nonsequential numbers). As expected both performance and retention were better after three than after two weeks of practice. Performance was better after a 1- week rather than a 4-week retention interval. Subjects with more original training had less performance loss.

Macek, Vilter, and Stubbs (1965) examined the effects of type of rehearsal on skill retention in a three-phase study. In the first phase and the first experiment, rehearsal was performed using verbal analogs: clock hours, calendar months, adjectives describing behavior, or no verbal analog. In the second experiment the same analogs were used but for six rather than 12 slots. In experiment 3, warm-up conditions were varied: cognitive warm-up, perceptual-motor warm-up, a combination of the first two, and no warm-up. The authors concluded from these three experiments that the greater the rehearsal relevancy, the smaller the initial retention decrement. Warm-up after two weeks was not as effective as weekly rehearsal over a six-week retention period. In the second phase, a single experiment was conducted using only four subjects. The independent variables were number of trials, retention interval, and level of experience. The results indicated that the effect of experience on initial retention performance was modest but resulted in markedly better performance after a short while. Continued warm-up reduced the difference between low and high experienced subjects. In phase three, warm-up without a visual display resulted in poorer performance overall than warm-up with a visual display regardless of the retention interval of 1, 4, or 5 weeks.

Using a different type of task, Melton (1964) examined the effects of target movement pattern (random versus nonrandom), display/control relationships (normal versus reverse), and retention interval (5 minutes, 1 day, 1 week) on tracking performance. As expected tracking performance was better in the normal than in the reverse condition and retention loss was greater for the reverse display. Unexpectedly, however, there was no effect of retention interval in the random target motion condition. This may have been due to insufficient training. Recovery from the retention loss occurred during the second or third retention trial. The subjects were 336 male undergraduate students. In a similar study, Swink, Trumbo, and Noble (1967) examined the effect of retention interval (3 versus 5 months) as well as task predictability (100 versus 75%), sequence length (number of targets 8 to 48), and training criteria (equal practice versus repetition). As expected, task predictability resulted in better performance. Their subjects were 120 male undergraduate students. Trumbo, Noble, Cross, and Ulrich (1964) also found a significant effect of predictability. In addition, there was a positive correlation between retention loss in a tracking task and retention interval (1 week, 1 month, or 5 months). A year later, Trumbo, Ulrich, and Noble (1965) reported that there was no effect of type of pretraining or display specificity after one month. In a later study, Trumbo, Noble, and Swink (1967) examined the effects of secondary task uncertainty in a series of three experiments. Their general conclusions were: 1) the performance of a secondary task decreases retention and 2) the decrement due to the secondary task is independent of the decrement produced by an 8-day retention interval.

Bernstein and Gonzalez (1971a) took a different approach. They asked their subject which types of training were most effective for a reaction-time task. The response was imagery was useful in learning. Their subjects were male undergraduate students. In a follow-on experiment, the authors reported that subjects (forty male undergraduates) trained with imagery outperformed those trained without imagery. The task was a reaction time task. Imagery had its greatest effect early in the test trials. Their findings were replicated in a series of basic psychology experiments reported in Bernstein and Gonzales (1971b). In still another approach, Boker (1974) investigated the effect of immediate versus delayed test time on reading performance. Not surprisingly, the subjects, 108 undergraduate students, retained more of the relevant than the incidental information.

Carron and Marteniuk (1970) measured the performance of 150 male high school students on a stabilometer after a 14-day retention interval. Subjects varied in balancing ability. After the 14-day retention interval, subjects with high balancing ability improved while low balancing ability subjects degraded. In a study looking at a longer retention interval, Carron (1971) reported that subjects with higher abilities on a stabilometer task retained their balancing performance better after two years than lower abilities subjects did although all groups did poorer over the two-year retention interval. Aptitude also had effects on performance on monitoring and rifle assembly tasks by Army recruits (Fox, Taylor, and Caylor, 1969). Not unexpectedly, recruits with lower aptitudes required more training. Grimsley (1969a) focused on training simulator fidelity. Sixty Army trainees received simulation on a missile launch control station in which everything worked, there was no electric power, or a reproduction. Unexpectedly, retention loss was equivalent over all three groups. Further, fidelity was unrelated to the amount of retraining necessary. In a related study Grimsley (1969b) examined the effect of method of retraining: instruction alone, instruction with the reproduction panel, or demonstration on the panel with the electrical power on. The retention intervals were 4 weeks, 4 weeks and then 2 more weeks, or 6 weeks. There was no difference between the 4- and 6-week retention intervals; however, the 6-week retention interval had fewer correct than the 4 and then 2-week retention interval. Using similar subjects, Vineberg (1975) measured performance of 200 Army soldiers on the Comprehensive Performance Test right after basic training and 6 weeks after basic training. Vineberg reported the average decrement over time was 18 to 26%.

Leonard, Wheaton, and Cohen (1976) also studied skill retention of Army personnel. They measured performance immediately after initial training, six weeks after initial training, seventeen weeks after initial training, six weeks after refresher training, and seventeen weeks after refresher training. Longer retention intervals were associated with poorer performance. Refresher training improved performance on some but not all tasks compared to those who did not have the refresher training.

In a more operational setting, Menglekoch, Adams, and Gainer (1960) examined the effects of amount of training (5 versus 10 trials) on instrument flying performance. There was greater retention loss on procedural tasks than for the tracking task. The loss was also greater for static or emergency procedures than for dynamic procedures. The performance of the more highly trained group was always superior to the less trained subjects. The number of training trials to attain the performance level on the final training after a retention interval of 120 days was greater for the group receiving more initial training in absolute but not in relative number of trials. The immediate performance after the retention interval was always superior in the greater trained group. The authors concluded that amount of training does count.

Caines and Danoff (1967) compared performance of military pilots who completed proficiency flying versus those who had flying duties. Their subjects included 84 F-4C pilots, 14 A-4E pilots, and 78 C-130E pilots. Pilots who had proficiency flying had more flying deficiencies than pilots who had flying duty. The deficiencies covered the complete range of takeoff, general air work, navigation, tactical information, bombing, tactics, and landing. Tasks associated with processing high rates of information while performing a simultaneous motor task had the greatest number of deficiencies. Further, pilots with two or more years of proficiency flying had significantly more deficiencies than those with less than two years of proficiency flying.

In nonmilitary flying, Hollister, La Pointe, Oman, and Tole (1973) examined the effect of recency of flight experience of private and commercial pilots on three test flights. They reported performance was positively related to total flight time and negatively related to years since certification as well as age. There were no effects of score on written quizzes, the subject’s own skill assessment, or the type of initial training received. The highest grades were received on those aspects of flight that were the most highly practiced: preflight and takeoff. The lowest scores were received on the infrequently practiced aspects such as stalls and instrument flight. In a similar study, Seltzer and McBrayer (1971) reported that performance of commercial pilots during a check ride declined continually until about 5.5 years after certification. The authors stated “This loss of proficiency is attributed partly to the motor skills of the individual pilot and also to his lack of knowledge”. The time for a commercial pilot to regain proficiency was 25 minutes of ground instruction and 1.5 hours of flight instruction. The time for private pilots was longer: 50 minutes ground instruction and 2.5 hours of flight. In the same year, Wilson (1973) examined the effect of prolonged non-flying periods on the pilot’s ability to perform a simulated carrier landing. For the study 15 naval aviators performed three carrier landings in a simulator. The aviators were current, one year stagnant, or two years stagnant. There was a small decrement between the current and one year groups but no difference between the one and two year groups. There was also a tendency for the aviators with more than 1 100 total flight hours to do better than aviators without as many flight hours.

Killian (1965) compared self-rated performance of second officers who had less than or more than 1,000 flight hours experience. All subjects had recently upgraded to first officers at United Air Lines. There were neither self-reported performance decrements nor any difference between high and low experienced pilots. Wright (1973) used a similar approach and had Army aviators complete a survey. The results suggest "flight excusal followed by refresher training would provide operational units with better qualified aviators at less cost than the traditional flying program" (p. 1).

In a different operational setting, Cotterman and Wood (1967) compared the performance of twelve test pilots performing a simulated lunar landing after 4, 8, 9, and 13 weeks. They reported that the longer the retention interval, the less probable that the landing would be successful. Youngling, Sharpe, Ricketson, and McGee (1968) examined the performance of simulated space missions as a function of retention period (30, 90, or 120 days). Their results indicate performance loss for the group that received less training (60 versus 120 trials) was twice that of the 120 trial group. There was a linear relationship between the length of the retention interval and the performance loss. Performance at the more difficult level (defined as performance tolerance) was retained better. Finally, reacquisition was more rapid after 30 than after 200 days.

In yet another operational setting, Johnson (1978) examined the effects of type of training and cognitive style on retention of a conveyor painting process that included 83 sequential steps and 24 numerical settings. There were three types of training:

1. Conventional practice – the trainee repeatedly performed the exact behavior as in the operational environment,
2. Reproduction practice – the trainee reproduced the control actions and system responses using pencil and paper, and
3. Blind practice – the trainee reproduced the same motor responses as in the conventional practice group but there was no visual stimulus.

The retention interval was 70 days. The subjects were paid individuals aged 16 to 34. There was no significant effect of the retention interval. Further, there was not a significant correlation between the number of errors on the last training trial and the retention scores. In a related study, Singer, Ridsdale, and Korienek (1979) examined the effects of learning strategy on acquisition, retention, and transfer of a visual tracking task. Strategies were use of imagery, rhythm, anticipation, informed choice, and none. The use of rhythm resulted in better performance than no strategy.

In yet another operational environment, Sitterley and Berge (1972) measured performance of space vehicle control from launch to orbit as well as emergency procedures over time (1 versus 6 months) and type of training (no practice, immediate rehearsal, distributed rehearsal, warm-up, and a combination of immediate and delayed rehearsal). In the no practice condition, there was a reliable decrement in altitude error at orbit insertion with time. The duration of the retention interval was not significantly related to the amount of degradation observed. In addition, performance improved with one warm-up practice and was at proficiency after five warm-up practices. These authors concluded, "In general, continuous control performance degradation was relatively moderate until 3 months had elapsed without practice. The data suggested that skill degradation had reached its peak at about 4 months" (p. 63). They added "Procedural performance, on the other hand, showed strong degradation after only 1 month without practice and a sharp increase in degradation at 4 months" (p. 63). In a follow-on experiment, Sitterley (1974) measured ability to land a space vehicle after four months. There were four types of training: no practice, static rehearsal, dynamic display, and self- paced static retraining. There were no crash landings for the group trained with static rehearsal and further this group had no significant skill degradation. These results are similar to an earlier study by Sitterley, Zaitzeff, and Berge (1972) on visual approach and landing tasks.

In an excellent review, Gardlin and Sitterley (1972) concluded from the research available at that time that "retention performance on a single specific task is apparently best for specific training; general training is superior when the amount of training increases and the task contains an element of uncertainty" (p. 4). Further, "test subjects typically reacquired their final training levels with the number of retraining trials fewer than 50 percent of the original training trials" (p. 15). They concluded, "It seems clear that the literature has identified the level of performance on the final training period as the primary predictor of skill retention for any given retention interval duration" (p. 20). In a later review, Smith and Matheny (1976) stated "Two major points supported by the literature and other evidence are cited: the first is that over learning of a task promotes its retention and the second is that motor skills will be retained longer than procedural or verbal materials" (p. 5).

In a similar review, Prophet (1976a) stated, "The single most important factor in determining absolute level of performance after periods of non-practice has consistently been found to be level of learning or skill prior to the nonpractice period" (p. 55). In addition, the "amount of decrement, i.e. the absolute amount forgotten, is largely independent of the level of initial skill or training and is much more a function of length of the nonpractice interval" (p. 55). Finally, the author stated, "The literature suggests that there is no fundamental difference between continuous control tasks and procedural tasks, as far as learning and retention are concerned if task organization is taken into account" (p. 62). This author also produced an excellent annotated bibliography (Prophet, 1976b). In a related conclusion, Hammerton (1963) reported that initial performance at the end of a six-month retention interval was significantly better for subjects receiving extensive training on a tracking task. In review focused only on motor skills, Schendel, Shields, and Katz (1978) concluded, "The single most important determinant of motor retention is level of original learning" (p. 1). They summarized that continuous control tasks are typically remembered for months or years. An example is riding a bicycle. The authors add that individuals with higher initial ability tend to achieve higher levels of proficiency and retain skill also at a higher level of proficiency.

Based on the above research, it is clear that practice of emergency procedures is especially critical as related to automated systems and skill retention. The first step in responding to an emergency is detecting that something went wrong. Research relevant to this area is discussed in the following section.

## 1.3 Is Something Wrong? - A Review of the Failure-Detection Research

The research reviewed in the section 5.1 dealt with observers actively searching for signals about objects outside the equipment. The typical scenario was that of the radar operator watching a scope for evidence of an enemy aircraft or ship. Signals were not malfunctions but rather reliable information about the world outside. The research to be reviewed in this section, however, deals with observers passively monitoring an automated system and searching for malfunctions in that system. The prime example is the pilot monitoring the autopilot. In this case the operator (i.e., the pilot) who has learned to rely on the automated system for crucial decision making must detect a failure in that system. Research examining this problem has been sparse.

An early study in this area was performed by Miller and Elkind (1967). They introduced failures (changes in the magnitude or polarity of gain) during performance of a compensatory tracking task. The authors reported that the detection latencies for these failures were highly variable, especially for gain increases and polarity reversals. In an earlier study, Miller (1965, cited in Young, 1969) reported that gain increases and reversals were quickly detected. He suggested that the short detection times associated with these failures were due to the large errors in the system that resulted when gain increases or polarity reversals went undetected. This suggests, not surprisingly, that the salience of the failure is a critical element in failure-detection performance. Rolfe (1972) that such performance is likely to occur:

1. "When an unexpected event is very probable.
2. When the operator is anxious.
3. When the operator's attention is being distracted.
4. During a period of reaction following a time of high stress.
5. When a set pattern of interpretation and action has been held for a long time" (p. 76).

Curry and Gai (1976) took a different approach and state that responding to a system failure can be decomposed into three functions: 1) failure detection, 2) failure identification, and 3) corrective action. Errors occur at any of these three functions when the operator's mental model of how the system works, based on his or her training and experience, does not match reality. Errors also occur whenever there is deficient information - either in quantity or accuracy. When the operator has deficient information, workload induced stress can cause deficiencies in cognitive functions. This is especially true for continuous failures rather than discrete failures and for decreases in frequency and variance in failures rather than for increases. The authors also hypothesized that fatigue and low motivation would result in even further deficiencies in cognitive functioning. Gail and Curry (1977) directly tested this hypothesis. Two pilots flew fully automated approaches and landings in an aircraft simulator. In several of the approaches, errors in the glide slope or airspeed indicators were introduced. The authors reported that failure-detection time decreased as failure magnitude increased. Further, the pilots tended to detect errors in the glide slope indicator faster than those in the airspeed indicator.

Vreuls, Barneby, Nichols, and Dent (1968) also examined the effect of failure salience on detection times. These researchers introduced three types of autopilot failure during final approaches flown in an aircraft simulator:

1. Passive control failure (dead fail),
2. Soft over control failure (0.25 degree/second drift), or
3. Flare mode engagement failure.

They reported that 39 of the 216 passive autopilot failures went undetected while only 5 of the 216 soft over autopilot failures were missed. The number of misses decreased when the pilots were provided with a caution light. Without a caution light, however, the detection latencies for passive failures (mean = 15.68 seconds) were reliably longer than those for soft over failure (mean = 7.15 seconds). The most difficult failure to detect (i.e., the one associated with the longest latencies) was a passive failure in the pitch axis. Vreuls, et al. also reported the occurrence of several inappropriate responses to system failures; pilots:

1. Misidentified the axis of the failure,
2. Disengaged the wrong axis, and
3. Disengaged the whole autopilot.

In another elaborate study, Ephrath and Curry (1977) measured the workload and failure-detection performance of 15 professional pilots landing an aircraft simulator in zero-zero conditions. The authors varied the parameters in Table 7. They found that workload (as measured by reaction time to a secondary, non-relevant stimulus onset) increased dramatically as the amount of automated control increased. Also, failure- detection times were longer when the failed axis was under manual rather than automatic control. Further, system failures went undetected only when the failed axis was being controlled manually and never when it was automatically controlled.

**Table 7 Variables Used by Ephrath and Curry (1977) in their Study of 15 Professional Pilots Landing on Aircraft Simulator**

| Type and Amount of Automated Control | Degree of Wind Disturbance | Type of System Failures |
|---|---|---|
| Completely automated flight control | No wind | Lateral axis failure |
| Manual control of lateral axis | | |
| Manual control of longitudinal axis | 45-degree Tailwind of 5 knots with Gusts to 15 knots | Longitudinal axis failure |
| Completely manual | 45-degree Tailwind of 10 knots with Gusts to 30 knots | No failure |

Other findings included that detection times for lateral-axis failures were longer than those for longitudinal failures at comparable workloads. Also, as expected, higher wind turbulence resulted in higher workloads and longer failure-detection times. However, failure-detection time was not a monotonic function of workload. Ephrath and Curry (1977) concluded from this that the participation mode (i.e., type and flight control) influenced failure detection independent of the associated differences in workload among modes.

From the findings described above, it would seem that automatic control is superior to manual control. However, other studies have not been as supportive. This is especially true of research performed by Wickens and Kessel at the University of Illinois. They have conducted experiments over a five year period that have consistently shown that operators generally are slower to detect a failure and recover from it when they are passively monitoring, rather than actively controlling, the system. Their research is described in detail below.

Kessel and Wickens (1978) reported two experiments that compared failure- detection performance in manual and automated modes. In the first experiment, step changes in system order were introduced into a 2-dimensional, pursuit-tracking task. Subjects were instructed to press a trigger on the joystick when they detected a failure. Kessel and Wickens found consistent superiority in failure-detection latency and accuracy when the operator was manually controlling the system rather than passively monitoring it. They attributed this difference to the proprioceptive cues available only in the manual control mode. In their second study, they directly tested this hypothesis by having subjects perform the tracking task using a joystick with 0 or 520 grams of resistance. Contrary to expectation, they found that the isotonic joystick did not degrade failure- detection performance.

Wickens and Kessel (1979) again compared failure-detection performance in manual and automatic control modes and again found longer failure-detection latencies in the automatic than the manual mode. Wickens and Kessel (1980) added a secondary task (either tracking or mental arithmetic) to their experimental paradigm and found that failure detection in the automatic mode was adversely affected by the mental-arithmetic task but not by the secondary tracking task. Conversely, failure detection in the manual mode was worse when the secondary task was tracking. Again, overall, failure-detection performance in the manual mode was characterized by shorter detection latencies and greater accuracy than that in the automatic mode.

Wickens and Kessel presented their conclusions in a general review paper published the next year.

> Our conclusions asserted that the impact of the loading tasks was upon the processing channels used to monitor the system, visual for the AU (automatic mode) detection, proprioceptive for MA (manual mode). The former utilizes resources associated with perception and central processing, the same resources as those demanded by the mental arithmetic task; the latter

> utilizes response-related resources, coincident with the subcritical loading task. A point of more general relevance here is that automation .does not necessarily eliminate or even reduce the workload of the human operator, but merely shifts the locus of processing demands (p. 146).

In a similar vein, Sheridan (1976) proposed the following general guideline for system control:

> In-the-loop participation (manual control) is best when error alone is insufficient or the input of motor 'identification signals' permits quick adaptation; but if control is advertently noisy or requires full attention to steering displays and leaves little time for other displays which offer important failure cues, the man should monitor and let the machine control (p. 176).

Rouse (1981) tried to resolve the automatic/manual control superiority controversy. He suggests that:

> It seems reasonable to conjecture that having to control while monitoring for failures is beneficial if performing the control task provides cues that directly help to detect failures and if the workload is low enough to allow the human to utilize these cues. Otherwise, controlling simply increases workload and decreases the amount of attention that can be devoted to failure detection (p. 83).

In comparing Wickens and Kessel's work (based on performing a two dimensional tracking task) with that of Curry, Ephrath, and Vreuls (landing aircraft), Rouse's suggestion seems eminently plausible.

Parasuraman, Molloy, Mouloua, and Hilburn (1996) also considered the effect of workload. These authors summarized a study in which subjects performed tracking and fuel-management tasks manually and either monitored an automated engine status task or performed this task manually. In the manual mode, subjects detected 70% of the engine malfunctions. In the automated mode, they detected less than 40%. For pilots, the detection rate was less than 60%. In another experiment, they reported that the detection rate of automation failures varied inversely with automation reliability.

In addition, Kessel (1986) performed four experiments using a simple collision avoidance game to assess the effects of automation on performance. She concluded that automation degraded operator performance because the operator had to predict and adapt to the automation system's actions. When the roles were reversed and the automated system had to adapt to operator performance, operator performance was enhanced.

Detection rates also seem to vary with pilot strategy. Beringer (1996) analyzed the recovery from automation malfunctions of 29 pilots none of whom had less than 300 hours of flight time. The data were collected in a fixed-base aircraft simulator. There were four types of failures: command over role (rate = 6 deg/sec), soft roll (rate = 1 deg/sec), soft pitch (rate = 0.2 deg/sec), and runaway pitch up trim. Response times were longer for covert changes (e.g., soft pitch ranged from 21.2 to 85.1 seconds) than for overt changes (e.g., command over roll ranged from 11.8 to 53.8 seconds). Two distinct response strategies were observed: immediate disconnect and manual override. In a follow up study, Beringer and Harris (1997) used the same simulator to observe malfunction recoveries of 24 pilots ranging in flight hours from 290 to 10,000 hours. Malfunction types were selected to cover the entire range from largely covert to largely overt: soft pitch failure rate = 0.2 deg/sec), attitude indicator failure (slow drift), runaway pitch down, and runaway roll servo (roll rate 12 to 15 degrees per second). First response and/or recognition times were commensurate with the overtness of the malfunction:

1. Soft pitch failure rate – 330 milliseconds to 73.7 seconds

2. Attitude indicator failure – 12.7 to 263 seconds

3. Runaway pitch down – 3.6 – 15.8
4. Runaway roll servo – 1.09 to 4.88 seconds.

The problem is not restricted to aviation but to any automated system. Kerstholt and Passenier (2000) reported the results of two experiments in maritime supervisory control. In the first experiment, 39 maritime-studies students diagnosed disturbances that could be real or false alarms. The presence of false alarms increased the rate at which the automated systems were sampled but also increased the problem solving time. The problem-solving time was longer in part because the students solved the problems sequentially not concurrently. This tunnel vision has been noted in many automation related accidents, e.g., Eastern L-1011, Miami Florida, December 29 1972. In the second experiment, the authors provided the student with either interactive support in which the student had to enter values related to the subsystems that were evidencing disturbances or a completely automated system that diagnosed the problem. The subjects were 30 maritime-studies students. Ironically even though both systems gave the same advice, more incorrect actions were taken with the completely automated system. When the support systems were removed, students who had used the interactive system showed a more structured approach to problem solving that student who had used the completely automated system.

Even with the controversial findings described above, there have even been attempts at developing a mathematical model of failure-detection performance. Specifically, Repperger, Haas, Schley, and Koivo (1998) have developed mathematical models to predict loss of control due to failures in the human interface device.

Summary

Automation has been used to enhance safety, increase productivity, reduce operator workload, and minimize error. But along with its many benefits have come problems. One of the most serious is a system failure that forces the operator (or user) to suddenly and unexpectedly enter the control loop. The operator, who has learned to rely on the automated system for crucial decision-making, must first detect the system failure and then take the steps necessary to recover. In some systems (e.g., aircraft, air traffic control, nuclear power plants) the operator must not only act, but also act quickly and confidently since recovery time may be critical.

To date, research comparing failure-detection performance in automated and manual control modes has been both sparse and contradictory. Work performed in the late seventies by Ephrath and Curry (1977) appeared to confirm the superiority of automatic over manual control in several areas including failure-detection performance. A series of studies by Kessel and Wickens (1978, 1982), however has repeatedly found longer failure-detection latencies in the automatic than in the manual mode. Other researchers (Ephrath, 1980; Stewart, 1978; Wiener and Curry, 1980b) have also reported problems associated with automated systems that range from no reduction in workload over manual systems to automation-induced errors. Clearly, additional research is necessary to discover the reasons for the contradictory findings and to establish the framework for making optimal use of the benefits inherent in automated systems.

Findings in the vigilance and failure-detection literature suggest the importance of operator arousal to performance. In the vigilance literature, stimulants, loud ambient noise, and the threat of electric shock improved detection performance. Lengthening signal durations, maintaining a high signal rate, increasing signal intensity, and increasing signal complexity also improved vigilance. It would seem that making the task more stimulating, improved performance.

In the failure-detection literature, performance was better in the automatic than in the manual mode when the control task was complex, i.e., piloting an aircraft (e.g., Ehprath and Curry, 1977). The converse was true, however, when the task was less complex, i.e., 2-D, pursuit-tracking (e.g. Kessel and Wickens, 1978).

### 1.4 Nothing Can Go Wrong –A Review of Automation-Induced Complacency Research

The ASRS defines complacency as "self-satisfaction, which may result in non- vigilance based on an unjustified assumption of satisfactory system state." Ropelski (1996) suggested that pilots suffer from complacency when they become too reliant on and confident of the automation. This can lead to accidents. He gives the example of the Air Inter A320 accident near Strasbourg France on January 20, 1992. The A320-pilot entered a trap involving low airspeed, low altitude, low power, and rising terrain into which the aircraft crashed without enough power to climb. System prevented pilot from raising the nose of the aircraft to avoid the terrain since this would have stalled the aircraft.

To help study automation complacency, Singh, Molloy, and Parasuraman (1992; 1993) developed and then validated a scale for measuring the potential for complacency. They tied the concept of complacency potential to "premature cognitive commitment". The scale has 20 items that are rated on a five-point Likert scale. 139 undergraduate students completed the original scale. Their responses were submitted to a factor analysis that identified the following factors: general automation, confidence, reliance, trust, and safety-related complacency. One example of an item on the scale is: "People save time by using automatic teller machines (ATMs) rather than a bank teller for banking transactions" (p. 23). An additional 175 undergraduates completed the scale. The results were similar.

In a more recent study, Prinzel, DeVries, Freeman, and Mikula (2001) examined the effects of three personal traits on performance in an automated task. The traits were:

1. Complacency potential as measured by the Complacency Potential Rating Scale (CPRS) in which subjects rate each of 20 items on a scale from 1 (strongly agree) to 5 (strongly disagree). This was the same scale as used by Singh, Molloy, and Parasuraman (1992). In addition, individuals were classified as high or low in complacency if they scored above or below the group median on this scale.

2. Boredom proneness as measured by the Boredom Proneness Scale (BPS) in which subjects answer each of 28 items with a yes or a no. A sample item is: "It is easy for me to concentrate on my activities".
3. Cognitive failure as measured by the Cognitive Failure Questionnaire (CFQ) in which subjects identify how often 25 items occurred to them in the last six months from very often (0) to very often (4). A sample item is: "Do you fail to notice sign posts in the road?"

The subjects were 40 undergraduate students. The tasks were monitoring an automated system for failures, managing fuel, and tracking. The last two tasks were performed manually. Subjects completed two 40-minute sessions. The sessions were scheduled one day apart. The results indicated that high complacency subjects did worse on both the system monitoring task and on the tracking task than those in the low complacency group. High complacency subjects also scored higher on the BPS than low complacency subjects. All three scales were significantly correlated: CPRS x BPS = +0.596, CPRS x CFQ = +0.709, and BPS x CFQ = +0.717. The authors concluded, "there are personality individual differences that are related to whether an individual will succumb to automation-induced complacency" (p. 35).

A related issue is over reliance on automation. Riley (1996b) proposed a theory to identify the factors associated with this phenomenon (see Figure 15). He then ran a series of simple computer-based experiments to test his theory. There were two tasks:
1) categorize a character as either a letter or a number and 2) correct random disturbances of a marker from a target location. Riley manipulated the reliability (50 versus 90% correct) of the automated system that could be selected to perform the categorization task. Workload and uncertainty were also varied. The dependent variable was the proportion of subjects who used automation. The overall use of automation was low: 35% under normal conditions to 50% in high workload. Further, there was no decrease in the proportion of subjects using automation after the automation failed. His subjects were undergraduate students. In a follow on experiment with pilots, one third of the pilots continued to use the automation throughout its failure period.

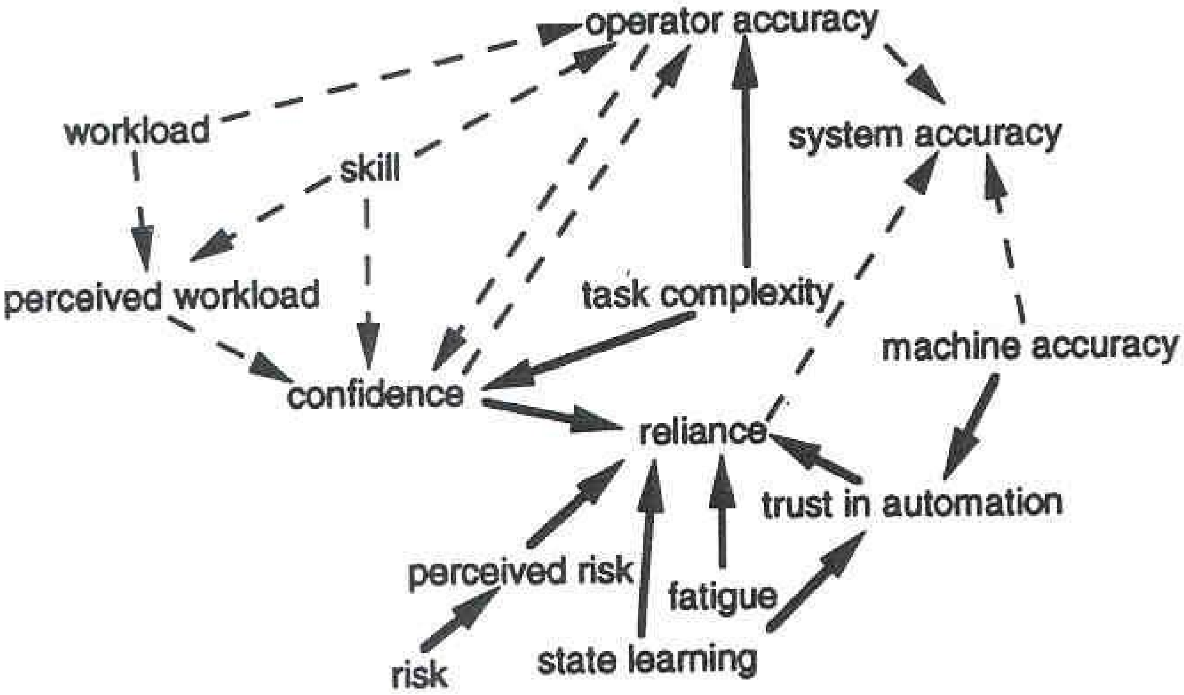


**Figure 15 Revised Theory Of Automation Use. Dotted Arrows Show Hypothesized Relationships that Have Not Been Confirmed by Experimental Evidence, Whereas Solid Lines Represent Those Relationships Supported From Evidence (Riley, 1996b, p. 33)**

1.5 The Knowing –Or Not –A Review of the SA Research

SA is knowledge relevant to the task being performed. For example, pilots must know the state of their aircraft, the environment through which they are flying, and relationships between them, such as thunderstorms are associated with turbulence. It is a critical component of decision-making and has been included in several models of decision-making (e.g., Dorfel and Distelmaier model, 1997; see Figure 16). SA has three levels (Endsley, 1991): Level 1, perception of the elements in the environment; Level 2, comprehension of the current situation; and Level 3, projection of future status. Endsley (1995) analyzed aviation accidents from 1989 to 1992 to identify SA problems. The results are presented in Table 8. As can be seen from the table, the predominant errors were related to vigilance and failure detection although errors occurred at all three levels of SA. Jones and Endsley (1996) applied the same methodology to ASRS incidents (see Figure 17). Clearly the major problem is failure to monitor. The authors went farther and explored why pilots failed to monitor automated systems. The results are presented in Figure 18. Not unexpectedly the most frequent reason was task distraction. This played a part in may of the cases presented in section 4 of this report. One example is the Eastern L-1011 accident near Miami Florida on December 29, 1972. This result directly impacts the design of automated systems.

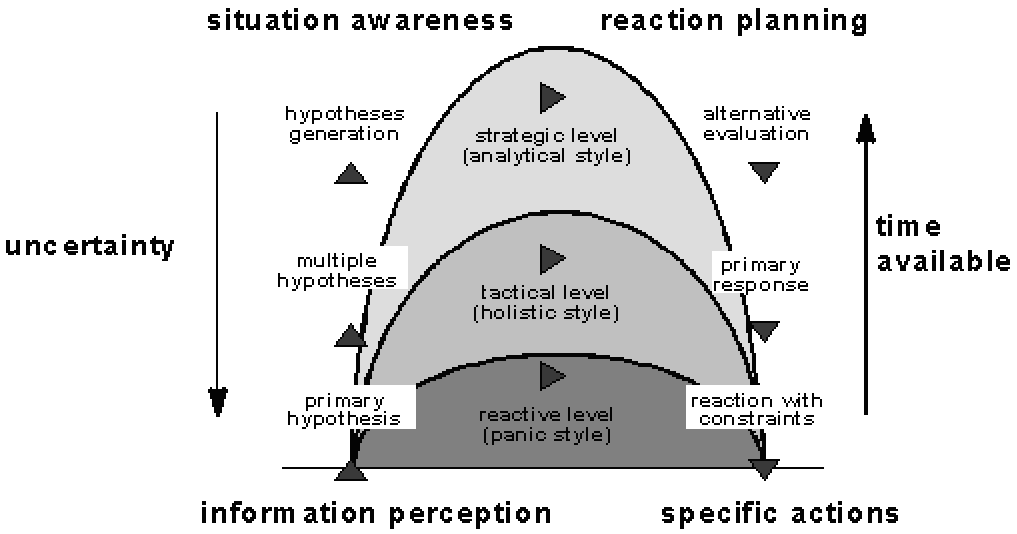


**Figure 16 Decision Making Under Uncertainty And Time Pressure (Dorfel and Distelmaier, 1997, p. 2)**

**Table 8 Aviation Accidents in Which SA was a Factor (Endsley, 1995, p.290)**

| Accident | Description | Position | Causal Factors | Other Factors |
|---|---|---|---|---|
| NTSB/AAR-92/05 | Spatial disorientation | Captain | Level 1 - misperception | Physiological |
| NTSB/AAR-92/01 | Crash on approach | Captain | Level 2 - other (significance) | Weather, Decision |
| NTSB/AAR-91/09 | Crash on take-off (icing) | Captain | Level 2 - no/poor model | Weather, Procedure, Physiological |
| NTSB/AAR-91/08 | Landing aircraft on occupied runway | Controller | Level 1 - difficult to detect, memory failure/task load, distraction | |
| NTSB/AAR-91/05 | Runway collision (fog) | Crew<br>Controller | Level 1 - difficult to detect<br>Level 1 - difficult to detect Level 2 - over-reliance on defaults | Weather, Procedure |
| NTSB/AAR-91/04 | Ran out of fuel | Crew<br>Controller | Level 2 - other (significance)<br>Level 3 - failure to project (time)<br>Level 1 - no data | Weather |
| NTSB/AAR-91/03 | Landed aircraft on occupied runway | Controller 1<br>Controller 2 | Level 2 - other (integration)<br>Level 1 - failure/distraction | |
| NTSB/AAR-91/01 | Loss of control - landing | First Officer | Level 1 - failure/omission | Psycho-motor |
| NTSB/AAR-90/05 | Crash into mountain | Crew | Level 1 - difficult to detect<br>Level 2 - wrong model | Weather, Decision Physiological |
| NTSB/AAR-90/04 | Struck power lines | Crew | Level 1 - misperception | Weather, Procedure |
| NTSB/AAR-90/03 | Crash on take- off<br>(Mistrimmed rudder) | Captain | Level 1 - failure/task load<br>Level 2 - wrong model | Procedure |

| | | | | |
|---|---|---|---|---|
| NTSB/AAR-90/02 | Loss of control (cargo door open) | First Officer<br>Captain<br>Captain | Level 1 - difficult to detect<br>Level 1 - no data (auto. Failure)<br>Level 3 - no/poor model | Mechanical<br>Psycho-motor |
| NTSB/AAR-89/04 | Crash on take- off<br>(Miss-set flaps & slats) | First Officer<br>Captain | Level 1 - failure/task load<br>Level 1 - no data (auto. Failure) | Procedure |
| NTSB/AAR-89/01 | Crash on approach | First Officer<br>Captain | Level 1 - failure/task load, misperception<br>Level 1 - failure/attn. Narrowing<br>Level 1 - failure/misperception | Decision |

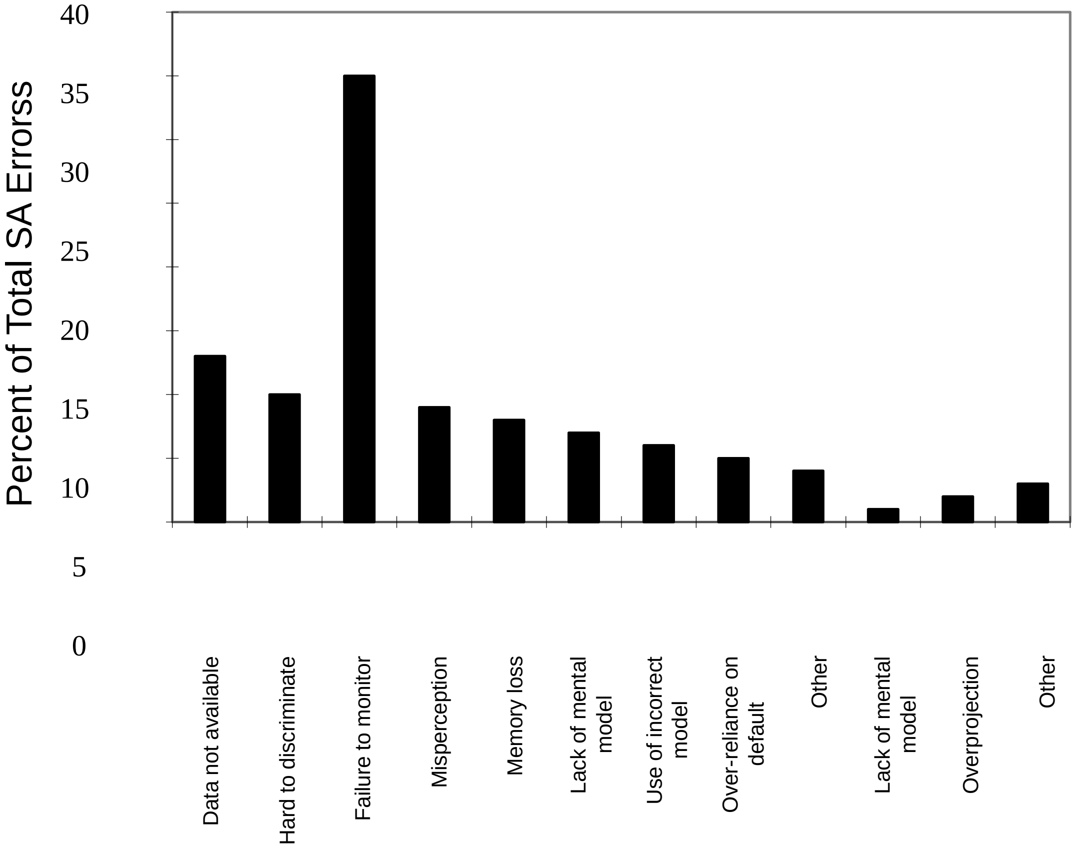


**Figure 17 Based on 143 ASRS Incidents January 1986-May 1992 (Jones and Endsley, 1996, p. 509)**

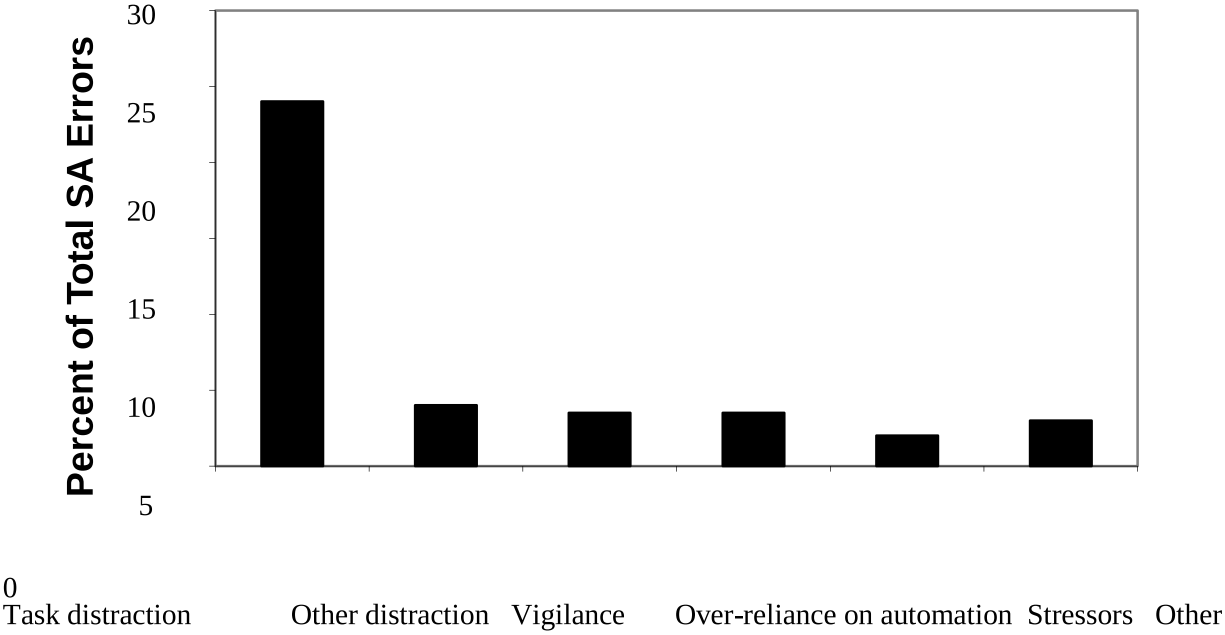


**Figure 18 SA Errors Associated with Failure to Monitor (Jones and Endsley, 1996, p. 510)**

In follow on work, Kaber and Endsley (1997) classified human supervisory control and monitoring of automated systems as forms of out-of-the-loop (OOTL) performance. As such they argue it is associated with the following negative consequences: “operator failure to observe system parameter changes and intervene when necessary (vigilance decrement), human over-trust in computers (complacency), operator loss of system of SA, and operator direct/manual control skill decay” (p. 126). The authors then suggest using LOA to counter the decrements in OOTL performance. Their LOA taxonomy is presented in Table 9.

**Table 9 LOA Taxonomy (Kaber and Endsley, 1997, p. 129)**

| LOA | Monitoring | Generating | Selecting | Implementing |
|---|---|---|---|---|
| 1. Manual control | Human | Human | Human | Human |
| 2. Action support | Human/computer | Human | Human | Human/computer |
| 3. Batch processing | Human/computer | Human | Human | Computer |
| 4. Shared control | Human/computer | Human/computer | Human | Human/computer |
| 5. Decision support | Human/computer | Human/computer | Human | Computer |
| 6. Blended decision making | Human/computer | Human/computer | Human/computer | Computer |
| 7. Rigid system | Human/computer | Computer | Human | Computer |
| 8. Automated decision making | Human/computer | Human/computer | Computer | Computer |
| 9. Supervisory control | Human/computer | Computer | Computer | Computer |
| 10. Full automation | Computer | Computer | Computer | Computer |

## 2. References

Abbott, K.H. and Schutte, P.C. Human-centered automation and AI: ideas, insights, and issues from the intelligent cockpit aids research effort. Presented at the workshop on Integrated Human Machine Intelligence International Joint Conference on Artificial Intelligence, 1989.

Adams, J.A. and Reynolds, B. Effect of shift in distribution of practice conditions following interpolated rest. Journal of Experimental Psychology, 1954, 47, 32-36.

Alluisi, E. Computer-aided Acquisition and Logistic Support - Human System Components (CALS-HSC). In E. Boyle, J. Ianni, J. Easterly, S. Harper, and M. Korna (Editors). Human-centered technology for maintainability: workshop proceedings (AL- TP-1991-0010). Armstrong Laboratory, Wright-Patterson Air Force Base, Ohio, June 1991.

Amalberti, R. Safety in process-control: an operator-centered point of view. Reliability Engineering and System Safety, 1992, 38, 99-108.

Ammons, R.B. Effect of distribution of practice on rotary pursuit “hits”. Journal of Experimental Psychology, 1951, 41, 17-22.

Ammons, R.B., Farr, R.P., Black, E., Neumann, E., Day, M., Marion, R., and Ammons, C.H. Long-term retention of perceptual motor skills. Journal of Experimental Psychology, 1958, 55, 318-328.

Badler, N.I. Human factors simulation research at the University of Pennsylvania. In E. Boyle, J. Ianni, J. Easterly, S. Harper, and M. Korna (Editors). Human-centered technology for maintainability: workshop proceedings (AL-TP-1991-0010). Armstrong Laboratory, Wright-Patterson Air Force Base, Ohio, June 1991.

Barnhart, W., Billings, C., Cooper, G., Gilstrap, R., Lauber, J., Orlady, H., Puskas, B., and Stephens, W. A method for the study of human factors in aircraft operations (NASA- TM-X-62472). Moffett Field, California: NASA Ames Research Center, September 1975.

Bennett, C.T. and Banks, W.W. The paradox of automation and the role of human centered design (UCRL-JC-112907). Livermore, California: Lawrence Livermore National Laboratory, 1993.

Beringer, D.B. Automation in General Aviation: Responses of pilots to autopilot and pitch trim malfunctions. Proceedings of the Human Factors and Ergonomics Society 40th Annual Meeting, 1996, 86 – 90.

Beringer, D.B. and Harris, H.C. Automation in General Aviation Part II: four ways to reach zero feet AGL unintentionally – autopilot and pitch trim malfunctions. Proceedings of the Human Factors and Ergonomics Society 41st Annual Meeting, 1997, 75 – 79.

Bernstein, B.R. and Gonzalez, B.K. Learning, retention, and transfer (12126-FR1-VOL- 1, NAVTRADEVCEN 0215-1-1 Volume I). Saint Paul, Minnesota: Honeywell, February 1971a.

Bernstein, B.R. and Gonzalez, B.K. Learning, retention, and transfer (NAVTRADEVCEN 68-C-0215-1-1 Volume II). Saint Paul, Minnesota: Honeywell, February 1971b.

Besco, R.O. and Funk, K. Conceptual design guidelines to rediscover systems engineering for automated flight decks. International Journal of Aviation Psychology, 1999, 9(2), 189-198.

Bhatt, S.C., Sun, B.K.H., and Thie, J. Human-centered automation of testing, surveillance, and maintenance. First Annual Instrumentation, Systems, and Automation Society/Electric Power Research Institute Joint Control and Automation Conference – Power Industry 34th Power Instrumentation Symposium, 1991, 489-499.

Billings, C.E. Aviation Automation: The search for a human-centered approach. Mahwah, New Jersey: Lawrence Erlbaum Associates, 1997.

Billings, C.E. Human-centered aircraft automation: A concept and guidelines (NASA-TM- 103885). Moffett Field, CA: National Aeronautics and Space Administration Ames Research Center, 1991a.

Billings, C.E. Human-centered aviation automation: principles and guidelines. NASA Technical Memorandum 110381, Moffett Field, California: NASA Ames, February 1996.

Billings, C.E. Toward a human-centered aircraft automation philosophy. The International Journal of Aviation Psychology, 1991b, 1 (4), 261-270.

Billings, C.E. Toward a human-centered aircraft automation philosophy. Proceedings of the 5th International Symposium on Aviation Psychology, 1989, 1-8.

Billings, C.E., Lauber, J.K., Cooper, G.E., and Ruffell-Smith, H.P. Retrospective studies of operating problems in air transport. Proceedings of Aircraft Safety and Operating Problems Conference at NASA Langley Research Center, 1976, 585-590.

Billingsley, J. and Schoenfisch, M. The successful development of a vision guidance system. Computers and Electronics in Agriculture, 1997, 16, 147-163.

Blakeslee, P. Attention and vigilance: performance and skin conductance response changes. Psychophysiology, 1979, 16 (5), 413-319.

Boehm-Davis, D.A., Curry, R.E., Wiener, E.L., and Harrison, R.L. Human factors of flight- deck automation – NASA/Industry Workshop (NASA-TM-81260). Moffett Field, California: NASA Ames Research Center, January 1981.

Boker, J.R. Immediate and delayed retention effects of interspersing questions in written instructional passages. Journal of Educational Psychology, 1974, 66, 96-98.

Bolman, L. Aviation accidents and the "theory of the situation". In G.E. Cooper (Editor) Resource Management on the Flight Deck (NASA Conference Publication 2120). Moffett Field, California: NASA Ames Research Center, 1979, 31.

Boyle, E. Human-centered technology: ends and means. In E. Boyle, S. Harper, J. Ianni, M. Korna, and J. Easterly Human centered technology for maintainability workshop proceedings.Wright-Patterson Air Force Base, Ohio: Armstrong Laboratory, June 1991.

Boyle, E., Easterly, J., and Ianni, J. Human-centered design for concurrent engineering. High Performance Systems, April, 1990, 58-60.

Boyle, E., Ianni, J., Easterly, J., Harper, S., and Korna, M. Human-centered technology for maintainability: workshop proceedings (AL-TP-1991-0010). Armstrong Laboratory, Wright-Patterson Air Force Base, Ohio, June 1991.

Braun, R.J. and Graeber, R.C. Human-centered designs in commercial transport aircraft. Proceedings of the Human Factors Society 36th Annual Meeting, 1992, 1118-1122.

Brinsfield, W.A. and Brown, R.G. Operator response to incidents. Transactions of the American Nuclear Society, 1989, 60, 198.

Broadbent, D.E. Decision and stress. New York, New York: Academic Press, 1971.

Brown, D.R., Briggs, G.E., and Naylor, J.C. The retention of discrete and continuous tasks as a function of interim practice with modified task requirements (AMRL-TDR-63- 35). Wright-Patterson Air Force Base, Ohio: Aerospace Medical Research Laboratory, May 1963.

Buckout, R., Naylor, J.C., and Briggs, G.E. Effects of modified task feedback during training on performance of a simulated attitude control task after 30 days (AMRL-TM- 63-125). Wright-Patterson Air Force Base, Ohio: Aerospace Medical Research Laboratory, December 1963.

Bureau of Air Safety Investigation Advanced technology aircraft safety survey report. http://www.atsb.gov.au/aviation/pdf/advtek.pdf

Caines, K.L.D. and Dainoff, M.N. Proficiency flying program study (PRC-R-952). Los Angeles, California: Planning Research Corporation, March 1967.

Carroll, J.M., Koenemann-Belliveau, J., Rosson, M.B., and Singley, M.K. Critical incidents and critical themes in empirical usability evaluation. Proceedings of the Eighth Human Computer Interface: People and Computers Conference, 1993, 279-292.

Carron, A.V. Effect of ability level upon retention of a balance skill after two years. Perceptual and Motor Skills, 1971, 33, 527-529.

Carron, A.V. and Marteniuk, R.G. Retention of a balance skill as a function of initial ability level. Research Quarterly, 1970, 41, 478-483.

Catalano, J.F. The effect of rest following massed practice of continuous and discrete motor tasks. Journal of Motor Behavior, 1978, 10, 63-67.

Chappell, S.L. The role of the pilot and the effects of automation: Lessons from incidents. 2.1-2.6.

Chiu, T. and Fon, D. Development of an automatic rolling system for rice seedlings. Journal of Agricultural Engineering Research, 2000, 76(2), 149-156.

Clark, L. Avionics incorporating human-centered design improves pilot interface with automated system. International Civil Aviation Organization Journal, May 1997, 11-12.

Corker, K.M., Cramer, M.L., and Henry, E.H. Methodology for evaluation of automation impacts on tactical command and control (C2) systems: implementation (AFHRL-TR- 90-9). Logistics and Human Factors Division, Air Force Human Resources Laboratory, Wright-Patterson Air Force Base, Ohio, June 1990.

Cotterman, T.E. and Wood, M.E. Retention of simulated lunar landing mission skill: A test of pilot reliability (AMRL-TR-66-222). Wright-Patterson Air Force Base, Ohio: Aerospace Medical Research Laboratories, April 1967.

Craig, A. Effect of prior knowledge of signal probabilities on vigilance performance of a two-signal task. Human Factors, 1980, 22 (3), 361-371.

Craig, A. Is the vigilance decrement simply a response adjustment towards probability matching? Human Factors, 1978, 20, 441-446.

Cugola, G., DiNitto, E., Fuggetta, A., and Ghezzi, C. A framework for formalizing inconsistencies and deviations in human-centered systems. Association of Computing Machinery Transactions on Software Engineering and Methodology, 5(3), 1996, 191- 230.

Curry, R.E. and Gai, E.G. Detection of random process failure by human monitors. In T.B. Sheridan and G. Johannsen (Editors) Monitoring behavior and supervisory control. New York, New York: Plenum, 1976.

Dekker, S.W.A. From meaningless toil to soporific monitoring: Seven lessons that perhaps were not applied to aviation automation after all or: the latent failure in the latent failure approach. Proceedings of the Eighth International Symposium on Aviation Psychology, 1995, 1, 181 – 185.

Dhiomasaigh, L.N. and McCarthy, J. Conflicts of accountability in aviation accidents and incidents. In D. Harris (Editor) International conference on engineering psychology and cognitive ergonomics. Aldershot, United Kingdom: Ashgate, 1, 323-329, 1997.

Dick, A.O., Bittner, A.C., Jr., Harris, R., and Wherry, R.J., Jr. Design of a MANPRINT tool for predicting personnel and training characteristics implied by system design (Army Research Institute Research Note 89-04). United States Army Research Institute for the Behavioral and Social Sciences, Alexandria, Virginia, January 1989.

Dieterly, D.L. John Henry syndrome: Man's conflict with automation. Proceedings of the 7th Symposium of Psychology in the Department of Defense, 1980, 301-308.

Doos, M., Backstrom, T., and Samuelsson, S. Evaluation of a strategy. Preventing accidents with automated machinery through targeted and comprehensive investigation conducted by safety engineers. Safety Science, 1994, 17, 187-206.

Dorfel, G. and Distelmaier, H. Enhancing situational awareness by knowledge-based user interfaces. In K. Garner (Editor). Proceedings for the Second Annual Symposium and Exhibition on Situational Awareness in the Tactical Air Environment. Patuxent River, MD: Naval Air Warfare Center Aircraft Division, 1997.

Dornheim, M.A. Autopilot malfunctioned while off in Miami incident. Aviation Week and Space Technology, 1 April 2001, 47-48.

Edan, Y. and Miles, G.E. Design of an agricultural robot for harvesting melons. Transactions of the American Society of Agricultural Engineers, 1993, 36(2), 593-603.

Edwards, E. Automation in civil transport aircraft. Applied Ergonomics, 1977, 8, 194- 198.

Edwards, E. Some aspects of automation in civil transport aircraft. In T.B. Sheridan and G. Johannsen (Editors) Monitoring behavior and supervisory control, New York, New York: Plenum, 1976.

Ekstrom, P.J. Analysis of pilot workloads in flight control systems with different degrees of automation. Paper presented at Institute of Radio Engineers International Congress, Human Factors Engineering in Electronics, Long Beach, California, 1962.

Endsley, M.R. A taxonomy of situation awareness errors in human factors in aviation operations. Proceedings of the 21st Conference of the European Association for Aviation Psychology, Vol. 3, Brookfield, Vermont: Avebury Aviation, 1995, 287-292.

Endsley, M.R. Situation Awareness in dynamic systems. In R.M. Taylor (Editor) Situational awareness in dynamic systems (IAM Report 708). Farnborough, United Kingdom: Royal Air Force Institute of Aviation Medicine, 1991.

Endsley, M.R. and Kaber, D.B. The impact of intermediate levels of automation on situation awareness and performance in dynamic control systems. Institute of Electrical and Electronics Engineers (IEEE) Sixth Annual Human Factors Meeting, 1997, 7-7 – 7- 12.

Endsley, M.R. and Kiris E.O. The out-of-the-loop performance problem and level of control in automation. Human Factors, 1995, 37(2), 381-394.

Endsley, M. and Strauch, B. Automation and Situation Awareness: the accident at Cali, Columbia. 877-881.

Ephrath, A.R. Pilot workload during an instrument approach. IEEE Transactions on Systems, Man, and Cybernetics, 1980, SMC-10(10), 676-678.

Ephrath, A.R. and Curry, R.E. Detection by pilots of system failures during instrument landings. IEEE Transactions on Systems, Man, and Cybernetics, 1977, SMC-7 (12), 841-848.

Evans, S.M. EDGE: a CAD tool for system design. In E. Boyle, J. Ianni, J. Easterly, S. Harper, and M. Korna (Editors). Human-centered technology for maintainability: workshop proceedings (AL-TP-1991-0010). Armstrong Laboratory, Wright-Patterson Air Force Base, Ohio, June 1991.

FAA Human Factors Team Report on interfaces between flight crews and modern flight deck systems. Washington, District of Columbia (DC): FAA, June 1996.

Feggetter, A.J. A method for investigating human factor aspects of aircraft accidents and incidents. Ergonomics, 1982, 25(11), 1065-1075.

Fisk, A.D. and Schneider, W. Controlled and automatic processing during tasks requiring sustained attention. A new approach to vigilance (TR 8006). Arlington, Virginia: Office of Naval Research, February 1980.

Fleischman, E.A. and Parker, J.F. Factors in the retention and relearning of perceptual- motor skills. Journal of Experimental Psychology, 1962, 64, 215-226.

Flight International Reverser blamed in Lauda crash report. Flight International, 1-7 September 1993, 5. Folds, D.J. Use of design reference scenarios in a user-centered design process. Presented at the 36th Aerospace Sciences Meeting and Exhibit, 1998, AIAA-98-1039.

Fox, W.L., Taylor, J.E., and Caylor, J.S. Aptitude level and the acquisition of skills and knowledge in a variety of military training tasks (TR 69-6). Alexandria, Virginia: Human Resources research Office, May 1969.

Frey, T.J. and Stults, J. Workstation prototypes a human engineering approach. Proceedings of the Human Factors Society 34th Annual Meeting, 1990, 547-551.

Funk, K. and Lyall, B. Human factors issues of flight deck automation. Proceedings of the 17th Digital Avionics Systems Conference, 1998, 1, E23/1-8.

Funk, K. and Lyall, B. Identifying flight deck automation issues (2000-01-5595). World Aviation Conference, 2000, 1-8.

Funk, K., Lyall, B., and Riley, V. Flight deck automation problems. Proceedings of the Eighth International Symposium on Aviation Psychology, 1995, 1, 265 - 269.

Funk, K., Lyall, B., Wilson, J., Vint, R., Niemczyk, M., Suroteguh, C., and Owen, G. Flight deck automation issues. International Journal of Aviation Psychology, 1999, 9(2), 109- 123.

Gaddy, D.E. Rig automation, dual activity alters work roles. Oil and Gas Journal, May 3, 1999, 69-78.

Gai, E.G. and Curry, R.E. Failure detection by pilots during automatic landing: Models and experiments. Journal of Aircraft, 1977, 14(2), 135-141.

Gardlin, G.R. and Sitterley, T.E. Degradation of learned skills. A Review and annotated bibliography (NASA-CR-128611). Washington, District of Columbia: Boeing, June 1972.

Garner, K.T. Proceedings for the second annual symposium and exhibition on Situational Awareness in the Tactical Air Environment. Patuxent River, Maryland: Naval Air Warfare Center Aircraft Division, June 1997.

George, L.J. and Mital, A. Case studies of implementation of safety devices in automation. Robotics and Autonomous Systems, 1989, 5, 173-181.

Glor, P.J. Human-centered design evaluation for enhanced system supportability. In E. Boyle, S. Harper, J. Ianni, M. Korna, and J. Easterly (Editors) Human centered technology for maintainability workshop proceedings. Wright-Patterson Air Force Base, Ohio: Armstrong Laboratory, June 1991.

Goodenough, D.J. and Metz, C.E. Effects of listening interval on auditory detection performance. Journal of Acoustical Society of America, 1974, 55(1), 111-116.

Graeber, C. and Billings, C.E. Human-centered automation: development of a philosophy. In S.A. Morello (Editor) Aviation Safety/Automation Program Conference (National Aeronautics and Space Administration Conference Publication 3090). Moffett Field, California: NASA Ames Research Center, October 1990, 91-104.

Grimsley, D.L. Acquisition, retention, and retraining: Effects of high and low fidelity in training devices (HUMRRO-TR 69-1). Alexandria, Virginia: George Washington University, Human Resources Research Office, February 1969a.

Grimsley, D.L. Acquisition, retention, and retraining: Group studies using low fidelity training devices (HUMRRO-TR 6904). Alexandria, Virginia: George Washington University, Human Resources Research Office, March 1969b.

Gubbels, A.W. and Craig, G.L. The NRC Bell 412 ASRA safety system: a human factors perspective on lessons learned from an airborne incident. Reliability Engineering and System Safety, 2002, 75, 273-283.

Haffar, I. Design and field evaluation of a low-cost crop transplanter with multiple seedlings feed. Agricultural Mechanization in Asia, Africa, and Latin America, 1995, 26(3), 29-32.

Hale, D.P., Sharpe, S., Haworth, D.A. Human-centered knowledge acquisition: a structural learning theory approach. International Journal of Human-Computer Studies, 1996, 45, 381-396.

Hammerton, M. Retention of learning in a difficult tracking task. Journal of Experimental Psychology, 1963, 66, 108-110.

Hancke, T. and Braune, R.J. Human-centered design of human-machine systems and examples from air transport. In G.C. Goodwin and R.J. Evans (Editors). Proceedings of the 12th Triennial World Congress of the International Federation of Automatic Control, 1994, 517-520.

Harlow, C. and Wang, Y. Automated accident detection system (01-2453). Transportation Research Record 1746, 2001, 90-93.

Harms-Ringdahl, L. Production disturbances – a safety problem in automatic machines. Man-Machine Systems, 1985, 231-236.

Harms-Ringdahl, L. Experiences from safety analysis of automatic equipment. Journal of Occupational Accidents, 1986, 8, 139-148.

Hart, F. Develop a field grid system for yield mapping and machine control Quarterly Report April 1 to June 30 1995 (DE95013820). Brookings, South Dakota: South Dakota State University, 1995.

Harwood, K. Defining human-centered system issues for verifying and validating Air Traffic Control Systems. Proceedings of the North Atlantic Treaty Organization Advanced Study Institute on Verification and Validation of Complex and Integrated Human-Machine Systems. 1993, 114-129.

Helmreich, R.L. Social psychology on the flight deck. National Aeronautics and Space Administration Conference Publication 2120, 1979, 17-30.

Henry, E.H. Designing Human-Centered Systems: Circa 2039 Scenario (AFHRL-TR-90- 32). Brooks Air Force Base, Texas: Air Force Human Resources Laboratory, August 1990.

Higgins, J.K. Boeing’s design approach for crew-centered flight decks. Proceedings of the Future Flight Deck – Safe and User Friendly, London, United Kingdom: Royal Aeronautical Society, 1996, 6.1-6.5.

Hollister, W.M., La Pointe, A., Oman, C.M., and Tole, J.R. Identifying and determining skill degradation of private and commercial pilots (FAA-RD-73-91). Cambridge, Massachusetts: Massachusetts Institute of Technology Measurement Systems Laboratory, June 1973.

Hughes, D. and Dornheim, M.A. Accidents direct focus on cockpit automation. Aviation Week and Space Technology, 1995, 142(5), 52.

Husseiny, A.A., Sabri, Z.A., Packer, D., Holmes, J.W., Adams, S.K., and Rodriguez, R.J. Operating procedure automation to enhance safety of nuclear power plants. Nuclear Engineering and Design, 1989, 277-297.

Hutchins, S.G. and Hutchins, R.G. User-centered interface design of tactical aircraft displays. 36th Aerospace Sciences Meeting and Exhibit, American Institute of Aeronautics and Astronautics, 98-1036, 1998.

Inagaki, T. Situation-adaptive responsibility allocation for human-centered automation. Transactions of the Society of Instrument and Control Engineers, 1995, 31(3), 292-298.

Inagaki, T., Takae, Y, and Moray, N. Decision support information for takeoff safety in human-centered automation: an experimental investigation of time-fragile characteristics. Proceedings of Institute of Electrical and Electronics Engineers Conference on Systems, Man, and Cybernetics, 1999, 1, 1101-1106.

Johnson, S.L. Retention and transfer of training on a procedural task; interaction of training strategy and cognitive style (Report Number DJ-60320M-1). Buffalo, New York: Calspan, January 1978.

Jones, D.G. and Endsley, M.R. Sources of situational awareness errors in aviation. Aviation, Space, and Environmental Medicine, 1996, 67(6), pp. 507-512.

Kaber, D.B. and Endsley, M.R. Out-of-the-loop performance problems and use of intermediate levels of automation for improved control system functioning and safety. Process Safety Progress, 1997, 16(3), 126-131.

Kaber, D.B., Riley, J.M., and Tan, K. Improved usability of aviation automation through direct manipulation and graphical user interface design. The International Journal of Aviation Psychology, 2002, 12(2), 153-178.

Kaplan, J.D. Synthesizing the effects of manpower, personnel, training, and human engineering. In E. Boyle, J. Ianni, J. Easterly, S. Harper, and M. Korna (Editors). Human-centered technology for maintainability: workshop proceedings (AL-TP-1991- 0010). Armstrong Laboratory, Wright-Patterson Air Force Base, Ohio, June 1991.

Karwowski, W., Rahimi, M., and Mihaly, T. Effects of computerized automation and robotics on safety performance of a manufacturing plant. Journal of Occupational Accidents, 1988, 10, 217-233.

Kennedy, R.S. and Coulter, X.B. Research note: The interactions among stress, vigilance, and task complexity. Human Factors, 1975, 17, 106-109.

Kerstholt, J.H. and Passenier, P.O. Fault management in supervisory control: the effect of false alarms and support. Ergonomics, 2000, 43(9), 1371-1389.

Kessel, C. and Wickens, C.D. Development and utilization of internal models in dynamic systems. A comparison of monitors and operators as failure detectors. Urbana- Champaign, Illinois: Engineering-Psychology Research Laboratory, Technical Report, ELP-78-2/AFOSR-78-5, December 1978.

Kessel, C.J. and Wickens, C.D. The transfer of failure-detection skills between- monitoring and controlling dynamic systems. Human Factors, 1982, 24(1), 49-60.

Kessel, K.L. Advanced aviation (AA) in tactical aircraft may have negative effects on pilot performance. Proceedings of the 1986 IEEE International Conference on Systems, Man, and Cybernetics, 1986, 1248-1253.

Killian, D.C. Pilot proficiency retention for United Air Lines Second Officers. United Air Lines, February 1965.

Kitazawa, M. and Sakai, Y. Human-centered automated vehicle system by PC networking. Proceedings of the Thirty-first International Manufacturing, Automation and Systems, Technology, Applications, Design, Organization and Management Research (MATADOR) Conference, 1995, 615-620.

Kjellen, U., Rundmo, T., Sandetorv, H., and Sten, T. Safety analysis of manual tasks in automatic production systems – implication for design. Accident Analysis and Prevention, 1990, 22(5), 475-486.

Kludze, A.K. Spacecraft mission operations – lessons learned from automation. 20th Digital Avionics Systems Conference, 2001, 1, 4.A.4-1 – 4.A.4 – 5.

Koshcheyev, L.A., Okin, A.A., and Moshkin, Y.A. AN adaptive centralized system of accident-prevention automation. Electrical Technology, 1991, 4, 1-12.

Kraiss, K. and Hamacher, N. Concepts of user centered automation. Aerospace Science and Technology , 2001, 5(8), 505-510.

Lagrange, V. An ergonomic perspective on automation. Institute of Electrical and Electronics Engineers International Workshop on Robot and Human Interactive Communication, 2001, 396-399.

Lang, R.P. and Vacroux, A.G. Operator-centered adaptive compensation of continuous manual control systems. Institute of Electrical and Electronics Engineers Transactions on Man-Machine Systems, MMS-10(2), 1969, 53-59.

Langle, T., Clavel, P., Goupillon, J., and Droulin, B. An ergonomics-oriented approach to design agricultural tractor's driver controls (972722). Society of Automotive Engineers, 1997.

Lauber, J.K. Resource management on the flight deck: Background and statement of the problem. Proceedings of a NASA/Industry Workshop on Resource Management on the Flight Deck, 1979, NASA Conference Publication 2120, 3-16.

Leonard, R.L., Wheaton, G.R., and Cohen, F.P. Transfer of training and skill retention (TR-76-A3). Washington, District of Columbia: American Institute for Research, October 1976.

Leverenz, F.L. and Chadwell, G.B. Considering human factors in process plant risk and hazard analyses. International Conference and Exposition for Advancing Measurement, Control and Automation Technologies, Products, and Services, 1999, 394, 89-96.

Liu, K.K. The highly automated airplane: Its impact on aviation safety and an analysis of training philosophy (AFIT/GSM/LAC/97J-1). San Antonio, Texas: Graduate School of Logistics and Acquisition Management, Air Force Institute of Technology, June 1997.

Lock, S., Allanson, J., and Phillips, P. User-driven design of a tangible awareness landscape. Designing Interactive Systems Processes Practices, Methods, and Techniques Conference Proceedings. New York, New York: Association for Computing Machinery, 2000, 434-440.

Macek, A.J., Vilter, P.F., and Stubbs, I.W. Rehearsal and warm-up in skill retention final report (Report Number 20153-FR-1). Minneapolis, Minnesota: Systems and Research Division, Honeywell, October 1965.

Martin, T., Kivinen, J., Rijnadorp, J.E., Rodd, M.G., and Rouse, W.B. Appropriate automation – integrating technical, human, organizational, economic, and cultural factors. Automatica, 27 (6), 901-907, 1991.

Melnick, M.J. Effects of over learning on the retention of a gross motor skill. Research Quarterly, 1971, 42, 60-69.

Melton, A.W. Retention of tracking skill final report (ORA Project 02855). Department of Psychology, University of Michigan, September 1964.

Mengelkoch, R.F., Adams, J.A., and Gainer, C.A. The forgetting of instrument flying skills as a function of the level of initial proficiency (NAVTRADEVCEN 71-16-18). Port Washington, New York: Naval Training Device Center, 1960.

Millard, D.L. Toward a reliable safety sensor implementation for industrial automation. International Journal of Industrial Ergonomics, 1991, 7, 277-286.

Miller, D.C. and Elkind, J.I. The adaptive response of the human controller to sudden changes in controlled process dynamics. IEEE Transactions on Human Factors in Electronics, 1967, HFE-8 (3), 218-223.

Mineo, Y., Suzuki, Y., Niinomi, T., Iwatani, K. and Sekiguichi, H. Safety assessment of factory automation systems. Electronics and Communications in Japan, 2000, Part 3, 83(2), 96-109.

Mitchell, C.M. Human-centered automation: a philosophy, some design tenets, and related research.

Human Interaction with Complex Systems: Conceptual Principles and Design Practice, 1996, 377-381.

Mitchell, C.M. Human-centered design of human-computer-human dialogs (NASA-CR- 197379). Atlanta, Georgia: Georgia Institute of Technology, July 1994.

Mitchell, C.M., Thurman, D.A., and Brann, D.M. The human factor in 'lights-out' automation: Using field study data to identify critical human factors design issues. Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting, 1998, 364 – 368.

Mitta, D.A. and Folds, D.J. Incident detection system design: the effects of high false alarm rate on operator performance. 1-7.

Mosier, K.L., Skitka, L.J., Heers, S.T., and Burdick, M.D. Automation bias, accountability, and verification behaviors. Proceedings of the 40th Annual Meeting of the Human Factors and Ergonomics Society, 1996, 1, 204-208.

National Transportation Safety Board Aircraft Accident/Incident Summary Report – Kansas City, Missouri April 13, 1987 (NTSB/AAR-89/01/SUM). Washington, DC: Bureau of Accident Investigation, March 31, 1989.

National Transportation Safety Board Aircraft Accident Report – Aeromexico DC-10-30, XA-DUH, Over Luxembourg, Europe, November 11, 1979 (NTSB-AAR-80-10). Washington, DC: Bureau of Accident Investigation, 7 November 1980.

National Transportation Safety Board Aircraft Accident Report – Air Transport International, Inc., Flight 805, Douglas DC-8-63, N794Alm Loss of control and crash Swanton, Ohio February 15, 1992 (NTSB?AAR-92/05). Washington, DC: National Transportation Safety Board, November 19, 1992.

National Transportation Safety Board Aircraft Accident Report – Aloha Islandair, Inc. Flight 1712 De Havilland Twin Otter, DHC-6-300, N707PV Halawa Point, Molokai,

Hawaii October 28, 1989 (NTSB/AAR-90/05). Washington, DC: Bureau of Accident Investigation, September 25, 1990.

National Transportation Safety Board Aircraft Accident Report – American Airlines Inc, DC-10-10, N110AA, Chicago-O'Hare International Airport, Chicago, Illinois May 25, 1979. (NTSB-AAR-79-17). Washington, DC: Bureau of Accident Investigation, December 21, 1979.

National Transportation Safety Board Aircraft Accident Report – Aviana, The Airline of Columbia, Boeing 707-321B, HK 2016 Fuel exhaustion Cove Neck, New York January 25, 1990 (NTSB/AAR-91/04). Washington, DC: Bureau of Accident Investigation, April 30, 1991.

National Transportation Safety Board Aircraft Accident Report – Collision with trees on final approach American Airlines Flight 1572 McDonnell Douglas MD-83, N566AA, East Granby, Connecticut November 12, 1995 (NTSB/AAR-96/05). Washington, DC: Bureau of Accident Investigation, November 13, 1996.

National Transportation Safety Board Aircraft Accident Report – China Airlines Boeing 747-SP, N4522V 300 Nautical Miles Northwest of San Francisco, California, February 19, 1985 (NTSB/AAR-86/03). Washington, DC: Bureau of Accident Investigation, March 29, 1986.

National Transportation Safety Board Aircraft Accident Report – Controlled Flight Into Terrain Korean Air Flight 801 Boeing 747-300, HL7468, Nimitz Hill, Guam, August 6, 1997 (NTSB/AAR-00/01). Washington, DC: National Transportation Safety Board, January 13, 2000.

National Transportation Safety Board Aircraft Accident Report – Crash During Landing Federal Express, Inc. McDonnell Douglas MD-11, N611FE, Newark International Airport, Newark, New

Jersey, July 31, 1997 (NTSB/AAR-00/02). Washington, DC: National Transportation Safety Board, July 25, 2000.

National Transportation Safety Board Aircraft Accident Report – Delta Air Lines, Inc. Boeing 727-232, N473DA Dallas-Fort Worth International Airport, Texas August 31, 1988 (NTSB/AAR-89/04). Washington, DC: National Transportation Safety Board, September 26, 1989.

National Transportation Safety Board Aircraft Accident Report – Descent Below Visual Glide path and Collision with Terrain Delta Air Lines Flight 554 McDonnell Douglas MD-88, N914DL LaGuardia Airport, New York October 19, 1996 (NTSB/AAR-97/03). Washington, DC: Bureau of Aviation Safety, 14 June 1973.

National Transportation Safety Board Aircraft Accident Report – Eastern Air Lines, Inc., L-1011, N310EA, Miami, Florida, December 29, 1972 (NTSB-AAR-73-14). Washington, DC: National Transportation Safety Board, August 25, 1997.

National Transportation Safety Board Aircraft Accident Report – Eastern Air Lines, Inc., L-1011, N334EA, Miami International Airport, Miami, Florida, May 5, 1983 (NTSB/AAR- 84/04). Washington, DC: National Transportation Safety Board, March 9, 1984.

National Transportation Safety Board Aircraft Accident Report– Evergreen International Airlines McDonnell Douglas DC-9-33F, N931F Saginaw, Texas March 18, 1989 (NTSB/AAR-91/02). Washington, DC: National Transportation Safety Board, April 23, 1990.

National Transportation Safety Board Aircraft Accident Report– Grand Canyon Airline Flight Canyon 5 De Havilland Twin Otter, DHS-6-300, N75GC Grand Canyon National Park Airport Tusayan, Arizona September 27, 1989 (NTSB/AAR-91/01). Washington, DC: National Transportation Safety Board, January 8, 1991.

National Transportation Safety Board Aircraft Accident Report– Inadvertent in-flight slate deployment China Eastern Airlines Flight 583 McDonnell Douglas MD-11, B-2171 950 Nautical Miles South of Shemya, Alaska April 6, 1993 (NTSB/AAR-93/07). Washington, DC: National Transportation Safety Board, October 27, 1993.

National Transportation Safety Board Aircraft Accident Report – In-flight Icing Encounter and Loss of Control Simmons Airlines, d.b.a American Eagle Flight 4184 Avions de Transport Regional (ATR) Model 72-212, N401AM Roselawn, Indiana October 31, 1994 Volume I (NTSB/AAR-96/01). Washington, DC: National Transportation Safety Board, July 9, 1996.

National Transportation Safety Board Aircraft Accident Report – In-flight Icing Encounter and Loss of Control Simmons Airlines, d.b.a American Eagle Flight 4184 Avions de Transport Regional (ATR) Model 72-212, N401AM Roselawn, Indiana October 31, 1994 Volume II (NTSB/AAR-96/02). Washington, DC: National Transportation Safety Board, July 9, 1996.

National Transportation Safety Board Aircraft Accident Report – L’Express Airlines, Inc., Flight 508, Beech C99, N7217L, Weather encounter and crash near Birmingham, Alabama July 10, 1991 (NTSB/AAR-92/01). Washington, DC: National Transportation Safety Board, March 3, 1992.

National Transportation Safety Board Aircraft Accident Report – Northwest Airlines, Inc., McDonnell Douglas DC-9-82, N312RC, Detroit Metropolitan Wayne County Airport, Romulus, Michigan, August 16, 1987 (NTSB/AAR-88/05). Washington, DC: National Transportation Safety Board, May 10, 1988.

National Transportation Safety Board Aircraft Accident Report – Northwest Airlines, Inc., Flights 1482 and 299 Runway incursion and collision Detroit Metropolitan/Wayne County Airport Romulus, Michigan December 3, 1990 (NTSB/AAR-91/05). Washington, DC: National Transportation Safety Board, May 10, 1988.

National Transportation Safety Board Aircraft Accident Report – Runway collision of Eastern Airlines Boeing 727, Flight 111 and Epps Air Service Beechcraft King Air A100 Atlanta Hartsfield International Airport Atlanta, Georgia January 18, 1990 (NTSB/AAR- 91/03). Washington, DC National Transportation Safety Board, May 29, 1991.

National Transportation Safety Board Aircraft Accident Report – Runway collision of USAir flight 1493, Boeing 737 and Skywest flight 5569 Fairchild Metroliner Los Angeles International Airport Los Angeles, California February 1, 1999 (NTSB/AAR-91/08). Washington, DC National Transportation Safety Board, October 22, 1991.

National Transportation Safety Board Aircraft Accident Report – Ryan International Airlines DC-9-15, N565PC Loss of control on takeoff Cleveland-Hopkins International Airport Cleveland, Ohio February 17, 1991 (NTSB/AAR-91/09). Washington, DC National Transportation Safety Board, November 16, 1991.

National Transportation Safety Board Aircraft Accident Report – Scandinavian Airlines System Flight 901, McDonnell Douglas DC-10-30, John F. Kennedy International Airport, Jamaica, New York, February 28, 1984 (NTSB/AAR-84/15). Washington, DC National Transportation Safety Board, November 15, 1984.

National Transportation Safety Board Aircraft Accident Report – Stall and Loss of Control on Final Approach Atlantic Coast Airlines, Inc., United Express Flight 6291 Jetstream 4101, N304UE, Columbus, Ohio, January 7, 1994 (NTSB/AAR-94/07). Washington, DC: National Transportation Safety Board, October 6, 1994.

National Transportation Safety Board Aircraft Accident Report – Trans-Colorado Airlines, Inc., Flight 2286 Fairchild Metro III, SA227 AC, N68TC Bayfield, Colorado January 19, 1988 (NTSB/AAR-89/01). Washington, DC: Bureau of Aviation Safety, 2 May 1973.

National Transportation Safety Board Aircraft Accident Report – Trans World Airlines, Inc., Boeing 707-313C, N788TW, John F. Kennedy International Airport, Jamaica, New York, December 12, 1972 (NTSB-AAR-73-11). Washington, DC: National Transportation Safety Board, February 4, 1989.

National Transportation Safety Board Aircraft Accident Report –USAir Boeing 737-400, LaGuardia Airport Flushing, New York September 20, 1989 (NTSB-AAR-90/03). Washington, DC: National Transportation Safety Board, July 3, 1990.

National Transportation Safety Board Aircraft Accident Report –USAir flight 105 Boeing 737-200, N283AU Kansas City International Airport, Missouri September 8, 1989 (NTSB-AAR-90/04). Washington, DC: Bureau of Aviation Safety, September 11, 1990.

National Transportation Safety Board Aircraft Accident Report –World Airways, Inc., DC- 8-63F, N802WA, King Cove, Alaska, September 8, 1973 (NTSB-AAR-74-6). Washington, DC: Bureau of Aviation Safety, 8 May 1974.

National Transportation Safety Board Aircraft Accident Report –World Airways, Inc., Flight 30H, McDonnell Douglas DC-10-30CF, N113WA, Boston-Logan International Airport, Boston, Massachusetts, January 23, 1982 (NTSB/AAR-85/06). Washington, DC: Bureau of Accident Investigation, 10 July 1985.

National Transportation Safety Board Marine Accident Report – Grounding of the Panamanian passenger ship Royal Majesty on Rose and Crown Shoal near Nantucket, Massachusetts June 10, 1995 (NTSB/MAR-97/01). Washington, DC: National Transportation Safety Board, April 2, 1997.

Naylor, J.C. and Briggs, G.E. Effective rehearsal of temporal and spatial aspects on long- term retention of a procedure skill. Journal of Applied Psychology, 1963, 47, 120-126.

Naylor, J.C., Briggs, G.E., and Reed, W.G. Task coherence, training time, and retention interval effects on skill retention. Journal of Applied Psychology, 1968, 52, 386-393.

Naylor, J.C., Briggs, G.E., and Reed, W.G. The effects of task organization on the retention of skill. Wright-Patterson Air Force Base, Ohio: Aerospace Medical Research Laboratory, September 1962.

Naylor, J.C., Briggs, G.E., Brown, E.R., and Reed, W.G. The effect of rehearsal on the retention of a time-shared task (AMRL-TDR-63-33). Wright-Patterson Air Force Base, Ohio: Aerospace Medical Research Laboratory, April 1963.

Nelson, W.R. and Haney, L.N. Development of a framework of human-centered automation for the nuclear industry. Idaho Falls, Idaho: Idaho National Engineering Laboratory, 1993.

Neumann, E. and Ammons, R.B. Acquisition and long-term retention of a simple, serial, perceptual-motor skill. Journal of Experimental Psychology, 1957, 53, 159-161.

Nordwall, B.D. Side-stick controllers give tactile feedback. Aviation Week and Space Technology, 15 Mary 2000.

Norman, D.A. Toward human-centered design. Technology Review, July 1993, 47-53.

North, D.W. Finding common ground in envelope protection systems. Aviation Week and Space Technology, 28 August 2000, 153(9), 66.

O'Hanlon, J.F. Research on the psychophysiological basis of human vigilance (TR 787- F). Arlington, Virginia: Office of Naval Research, April 1979.

Papastavrou, J.D. and Lehto, M.R. Improving the effectiveness of warning by increasing the appropriateness of their information content: some hypotheses about human compliance. Safety Science, 1996, 21, 175-189.

Parasuraman, R. Memory load and event rate control sensitivity decrements in sustained attention (TR-26). Arlington, Virginia: Office of Naval Research, February 1980.

Parasuraman, R., Molloy, R., Mouloua, M., and Hilburn, B. Monitoring of automated systems. In R. Parasuraman and M. Mouloua (Editors). Automation and human performance: theory and applications. Mahwah, New Jersey: Lawrence Erlbaum Associates, 1996, 91-115.

Parasuraman, R. and Riley, V. Humans and automation: Use, misuse, disuse, abuse. Human Factors, 1997, 39(2), 230 – 253.

Parasuraman, R., Hilburn, B., Molloy, R., and Singh, I. Adaptive automation and human performance: III. Effects of practice on the benefits and costs of automation shifts (NAWCADWAR-92037-60). Warminster, Pennsylvania: Naval Air Warfare Center, October 1991.

Park, K.S. and Han, S.W. Performance obstacles in cellular manufacturing implementation – empirical investigation. Human Factors and Ergonomics in Manufacturing, 2002, 12(1), 17-29.

Pearson, C.H. and Robinson, L. Automating a commercial swather for harvesting forage plots. Agronomy Journal, 1994, 86, 1131-1133.

Pedersen, J.F. and Moore, K.J. An automated plot harvest system for use with a commercial forage harvester. Agronomy Journal, 1995, 87, 605-607.

Peel, G., Michielen, M., and Parker, G. Some aspects of road sweeping vehicle automation. Institute of Electrical and Electronics Engineers /American Society of Mechanical Engineers International Conference on Advanced Intelligent Mechatronics Proceedings, 2001, 337-342.

Phillips, E.H. Survey revealed automated cockpits harbor safety pitfalls. Aviation Week and Space

Technology, 22 November 1999, 151(21), 48.

Poole, R.A., Golde, P.V., Baiden, G.R., and Scoble, M. A review of INCO's mining automation efforts in the Sudbury Basin. Canadian Mining and Metallurgical Bulletin, 1998, 91(1016), 68-74.

Poulton, E.C. The effect of fatigue upon inspection work. Applied Ergonomics, 1973, 4, 73-83.

Prinzel, L.J., DeVries, H., Freeman, F.G., and Mikulka, P. Examination of automation- induced complacency and individual difference variates (NASA/TM-2001-211413). Hampton, Virginia: National Aeronautics and Space Administration, December 2001.

Prophet, W.W. Long-term retention of flying skills: A review of the literature (HumRRO- FR-ED(P)-76-35). Alexandria, Virginia: Human Resources Research Organization, October 1976a.

Prophet, W.W. Long-term retention of flying skills: An annotated bibliography (HumRRO- FR-ED(P)-76-36). Alexandria, Virginia: Human Resources Research Organization, October 1976b.

Repperger, D.W., Haas, M.W., Schley, P.C., and Koivo, A.J. Failure detection methods to predict loss of control involving human-interface devices Part I: Theory. Proceedings of the American Control Conference, 1998, 2862-2866.

Reynard, W.D Incident reporting – Its role in aviation safety and the acquisition of human error data (ADP002699). Moffett Field, California: NASA Ames Research Center, October 1983.

Richter, J.J., Barnett, D.S., and Alluisi, E.A. Research and development strategies for human centered and group support technologies (IDA Paper P-2630). Washington, District of Columbia: Office of the Under Secretary of Defense, May 1992.

Riley, V. Human-centered design and the revolutionary flight deck: turning pilots back into pilots. Proceedings of the Future Flight Deck – Safe and User Friendly, London. United Kingdom: Royal Aeronautical Society, 1996a, 11.1-115.

Riley, V. Operator reliance on automation: Theory and data. In R. Parasuraman and M. Mouloua (Editors). Automation and human performance: theory and applications. Mahwah, New Jersey: Lawrence Erlbaum Associates, 1996b, 19-36.

Riley, V. What avionics engineers should know about pilots and automation. Institute of Electrical and Electronics Engineers, 1995, 252-257.

Roberts, D.P. and Gawron, V.J. Operator/Manager Classification as a Precursor to Function Allocation. Proceedings of the 30th Annual Meeting of the Human Factors Society, 1986, 857-860.

Roebuck, J.A., Jr. Overcoming barriers to computer human modeling in concurrent engineering. In E. Boyle, J. Ianni, J. Easterly, S. Harper, and M. Korna (Editors). Human-centered technology for maintainability: workshop proceedings (AL-TP-1991- 0010). Armstrong Laboratory, Wright-Patterson Air Force Base, Ohio, June 1991.

Rogers, W.H. What implications does human-centered automation have for situational awareness requirements. Proceedings of the Human Factors Society 34th Annual Meeting, 1990, 23.

Rolfe, J.M. Ergonomics and air safety. Applied Ergonomics, June 1972, 75-81. Ropelewski, R.

Control in the cockpit. Aerospace America, August 1996, 28-33.

Roske-Hofstrand, R. J. Video in applied cognitive research for human-centered design. Special Interest Group for Computer-Human Interaction Bulletin, October 1989, 21 (2), 75-77.

Rouis, D. Automated planning for restarting batch procedure in the wake of an incident (single product). Proceedings of the Institute of Electrical and Electronics Engineers/Economic Commission for Latin America/International Federation for Information Processing International Conference on Architectures and Design Methods for Balanced Automation Systems, 1995, 75-87.

Rouse, W.B. Design for success A human-centered approach to designing successful products and systems. New York: John Wiley & Sons, Incorporated, 1991.

Rouse, W.B. Human-computer interaction in the control of dynamic systems. Computing Surveys, 1981, 13 (1), 71-99.

Rudisill, M. Crew/automation interaction in space transportation systems: Lessons learned from the Glass Cockpit. Proceedings of the Human Space Transportation and Exploration Workshop, 2000.

Sarter, N.B. Cockpit automation: from quantity to quality, from individual pilot to multiple agents. In R. Parasuraman and M. Mouloua (Editors). Automation and human performance: theory and applications. Mahwah, New Jersey: Lawrence Erlbaum Associates, 1996, 267-280.

Sarter, N.B. and Woods, D.D. From tool to agent: the evolution of (cockpit) automation and its impact on human-machine coordination. Proceedings of the 39th Annual Meeting of the Human Factors and Ergonomics Society. 1995, 1, 79-83.

Scerbo, M.W. Adaptive automation: Who can you turn to when things heat up (AIAA-99- 3503)? Proceedings of the 33rd Thermophysics Conference, American Institute of Aeronautics and Astronautics, 1999.

Schendel, J.D., Shields, J.L., and Katz, M.S. Retention of motor skills: review. Alexandria, Virginia: Army research Institute, September 1978.

Schiro, C. Fieldbus for task farm automation. Control, February 2002, 42-45.

Schroer, R. Fly-by-wire to drive-by-wire: are automated highways straight ahead? Institute of Electrical and Electronics Engineers Aerospace and Electronic Systems Magazine, September 2001, 16(9), 2.

Schutte, P.C. and Willshire, K.F. Designing to control flight crew errors. Institute of Electrical and Electronics Engineers International Conference on Systems, Man, and Cybernetics. New York, New York: Institute of Electrical and Electronic Engineers, 1997, pp. 1978-1983.

Seltzer, L.Z. and McBrayer, J.D. A study of the effect of time on the instrumented skill of the private and commercial pilots (FAA-DS-70-12). Cahokia, Illinois: Parks College of Aeronautical Technology of Saint Louis University, March 1971

Sheridian, T.B. Preview of models of the human monitor/supervisor. In T.B. Sheridan and G. Johannsen (Editors) Monitoring behavior and supervisory control. New York: Plenum, 1976.

Singer, R.N., Ridsdale, S., and Korienek, G.G. The influence of learning strategies in the acquisition, retention, and transfer of a visual tracking task (ARI TR 402). Tallahassee, Florida: Florida State University, August 1979.

Singh, I.L., Molloy, R., and Parasuraman, R. Automation-induced “complacency”: Development of the Complacency-Potential Rating Scale. International Journal of Aviation Psychology, 1993, 3(2), 111-122.

Singh, I.L., Molloy, R., and Parasuraman, R. Development and validation of a scale of automation-induced “complacency”. Proceedings of the Human Factors Society 36th Annual Meeting, 1992, 22 – 25.

Sitterley, T.E. Degradation of learned skills – static practice effectiveness for visual approach and landing skill retention (NASA-CR-140225). Seattle, Washington: Boeing, May 1974.

Sitterley, T.E. and Berge, W.A. Degradation of learned skills (NASA CR 128612). Effectiveness of practice methods on simulated space flight skill retention. Seattle, Washington: Boeing, July 1972.

Sitterley, T.E., Zaitzeff, L.P., and Berge, W.A. Degradation of learned skills. Effectiveness of practice methods on visual approach and landing skill retention (NASA- CP-128912). Seattle, Washington: Boeing, October 1972.

Small, R.L. and Rouse, W.B. Certify for success: A methodology for human centered certification of advanced aviation systems. In J.A. Wise, V.D. Hopkin, and D.J. Garland (Editors) Human factors certification of advanced aviation technologies (NASA-CR- 199009). Daytona Beach, Florida: Embry-Riddle Aeronautical University Press, 1994.

Smith, B.R. Six years into the A3I program: progress and problems. In E. Boyle, J. Ianni,
J. Easterly, S. Harper, and M. Korna (Editors). Human-centered technology for maintainability: workshop proceedings (AL-TP-1991-0010). Armstrong Laboratory,
Wright-Patterson Air Force Base, Ohio, June 1991.

Smith, D. and Dieterly, D.L. Automation literature: A brief review and analysis (NASA- TM-81245). Moffett Field, CA: National Aeronautics and Space Administration Ames Research Center, October 1980.

Smith, J.F. and Matheny, W.G. Continuation versus recurrent pilot training (AFHRL-TR- 76-4). Brooks Air Force Base, Texas: Air Force Human Resources Laboratory, May 1976.

Smith, P., Woods, D., McCoy, E., Billings, C., Sarter, N., Denning, R., and Dekker, S. Human-centered technologies and procedures for future air traffic management (NASA- CR 205087). Columbus, Ohio: The Ohio State University Research Foundation; January 1997.

Sochet, E. and Lemanski, R. Automated flight plan filing by simulated voice recognition (FAA-RD-80-56). Atlantic City, New Jersey: Federal Aviation Administration Technical Center, August 1980.

Sogame, H. and Ladkin, P. Aircraft Accident Investigation Report 96-5.

Spady, A.A. Preliminary report on airline pilot scan patterns during simulated ILS approaches. Proceedings of Aircraft Safety and Operating Problems Conference at NASA Langley Research Center, 1976, 603-616.

Sparaco, P. Incident prompts Airbus to alter A320 AOA limits. Aviation Week and Space Technology, 2001, 154(22), 38.

Spector, A. and Gifford, D. The space shuttle primary computer system. Communications of the ACM, 1984, 27(9), 872-900.

Spelt, P.F. and Jones, S.L. Operator-centered control of a semi-autonomous industrial robot. Oak Ridge, Tennessee: Oak Ridge National Laboratory, 1994a.

Spelt, P.F. and Jones, S.L. Operator-centered control of a semi-autonomous industrial telerobot. Proceedings of the Human Factors and Ergonomics Society 38th Annual Meeting, 1994b, 1048-1051.

Stanton, N.A. and Marsden, P. From fly-by-wire to drive-by-wire; safety implications of automation in vehicles. Safety Science, 1996, 24(1), 35-49.

Stern, J.A. and Wang, L. Automation and automobile accidents. 28th International Symposium on Automotive Technology and Automation Proceedings for the Dedicated Conference on Robotics Motion and Machine Vision in the Automotive Industries, 1995, 449-457.

Stewart, S.R. Utility of automation of order of battle and target intelligence data for intelligence analysis. Alexandria, Virginia: U.S. Army Research Institute for the Behavioral and Social Sciences, ARI-RR-1194, September 1978.

Stolze, M., Gutknecht, M., and Pfeifer, R. Building human-centered intelligent cooperative information systems with IKEA. In M.P. Papazoglou and J. Zeleznikow (Editors). Next generation of information systems. Berlin: Springer-Verlag, 1992.

Strauch, B. Automation and decision-making – lessons learned from the Cali accident. Proceedings of the Human Factors and Ergonomics Society 41st Annual Meeting, 1997, 195-199.

Summers, A. The safe automation gap – Part 1. Hydrocarbon Processing, April 2002, 85-87.

Summers, A. The safe automation gap – Part 2. Hydrocarbon Processing, May 2002, 57-60.

Swets, J.A. (Editor) Signal detection and recognition by human observers. New York: Wiley, 1964.

Swets, J.A., Tanner, W.P., and Birdsall, T.G. Decision processes in perception. Psychological Review, 1961, 68, 201-340.

Swink, J., Trumbo, D., and Noble, M. On the length-difficulty relation in skill performance. Journal of Experimental Psychology, 1967, 74, 356-362.

Tetmeyer, D.C. Human-centered focus in systems engineering. In E. Boyle, J. Ianni, J. Easterly, S. Harper, and M. Korna (Editors). Human-centered technology for maintainability: workshop proceedings (AL-TP-1991-0010).Armstrong Laboratory, Wright-Patterson Air Force Base, Ohio, June 1991.

Thackray, R.I. Boredom and monotony as a consequence of automation: A consideration of the evidence relating boredom and monotony to stress (FAA-AM-80-1). Oklahoma City, Oklahoma: Federal Aviation Administration Aeromedical Institute, February 1980.

Thackray, R.I., Bailey, J.P., and Touchstone, R.M. The effect of increases monitoring load on vigilance performance using a simulated radar display. Ergonomics, 1979, 22 (5), 529-539.

Thein, B. Human performance modeling: an integrated approach. In E. Boyle, J. Ianni, J. Easterly, S. Harper, and M. Korna (Editors).Human-centered technology for maintainability: workshop proceedings (AL-TP-1991-0010). Armstrong Laboratory, Wright-Patterson Air Force Base, Ohio, June 1991.

Thompson, C.P., Wenger, S.K., and Bartling, C.A. How recall facilitates subsequent recall: A reappraisal. Journal of Experimental Psychology: Human Learning and Memory, 1978, 4(3), 210-221.

Toola, A. Elicitation of safety requirements for process automation. Reliability Engineering and System Safety, 1992, 35, 209-215.

Toola, A. The safety of process automation. Automatica, 1993, 29(2), 541-548.

Trumbo, D., Nobel, M., and Swink, J. Secondary task interference in the performance of tracking tasks. Journal of Experimental Psychology, 1967, 73, 232-240.

Trumbo, D., Noble, M., Cross, K., and Ulrich, L. Task predictability in the organization, acquisition, and retention of tracking skills. Journal of Experimental Psychology, 1964, 70, 252-263.

Trumbo, D., Ulrich, L., and Noble, M.E. Verbal coding and display coding in the acquisition and retention of tracking skill. Journal of Applied Psychology, 1965, 49, 368- 375.

Tsoukalas, L.H. Virtual measurement and prediction in human-centered automation. Proceedings of the Topical Meeting on Computer-Based Human Support Systems: Technology, Methods, and Future, 1995, 235-241.

United Nations The human consequences of the Chernobyl nuclear accident – A strategy fir recovery. http://www.undp.org/dpa/publications/chernobyl.pdf, January 25, 2002.

Vakil, S.S. and Hansman, R.J. Analysis of complexity evolution management and human performance issues in commercial aircraft automation systems (ICAT-2000-3). Cambridge, Massachusetts: International Center for Air Transportation, Massachusetts Institute of Technology, May 2000.

Van Dusen, F. and Schlosberg, H. Further study of the retention of verbal and motor skills. Journal of Experimental Psychology, 1948, 38, 526-534.

Vineberg, R. A study of the retention of skills and knowledge acquired in basic training (HUMRRO-TR-75-10). Alexandria, Virginia: Human Resources Research Laboratory, June 1975.

Vreuls, D., Barneby, S.F., Nichols, D.E., and Dent, P.L. Pilot failure detection performance with three levels of fault warning information. System Research on Development Service, FAA, SRDS RD 68-9, February 1968.

Wainwright, W. Integration of Situational Awareness on the Airbus flight decks. 4.1-4.9. Walker, J.A., Ruberg, G.E., and O'Dell, J.J. Simulation for emergency management: Taking advantage of automation in emergency preparedness. Simulation, September 1989, 95-100.

Warm, J.S., Dember, W.N., and Hancock, P.A. Vigilance and workload in automated systems. In R. Parasuraman and M. Mouloua (Editors). Automation and human performance: theory and applications. Mahwah, New Jersey: Lawrence Erlbaum Associates, 1996, 183-200.

Warren, N.D. Automation, human engineering, and psychology. American Psychologist, 1956, 11, 531-536.

Wegner, K.W. A tentative taxonomy of human interactive factors in aircraft mishaps (AFOSR-TR-80-0595). Chestnut Hill, Massachusetts: Boston College, May 1980.

Welford, A.T. Signal, noise, performance, and age. Human Factors, 1981, 23(1), 97- 109.

Whitaker, R. Fuel savers in the cockpit. Flight International, 28 February 1981, 563-568. Wickens, C.D.

and Kessel, C. Failure detection in dynamic systems. In J. Rasussen and
W.B. Rouse (Editors), Human detection and diagnosis of system failures. New York:
Plenum Press, 1981.

Wickens, C.D. and Kessel, C. Processing resource demands of failure detection in dynamic systems. Journal of Experimental Psychology: Human Perception and Performance, 1980, 6, 564-577.

Wickens, C.D. and Kessel, C. The effects of participatory mode and task workload on the detection of dynamic system failures. IEEE Transactions on Systems, Man, and Cybernetics, 1979, SMC-9, 24-34.

Wiener, E.L. Fallible humans and vulnerable systems: lessons learned from automation. In J.A. Wise and A. Debons, (Eds.). Information systems: failure analysis. Berlin: Springer-Verlay, 1987, 163-181.

Wiener, E.L. Midair collision: The accidents, the systems, and the realpolitik. Human Factors, 1980, 22, 521-533.

Wiener, E.I. and Curry, R.E. Automation in the cockpit. Proceedings of the Human Factors Society 24th Annual Meeting, 1980a, 38.

Wiener, E.L. and Curry, R.E. Flight-deck automation: Promises and problems. Ergonomics, 1980b, 23, 995-1011.

Wiener, E.L. and Curry, R.E. Flight-deck automation: Promises and problems (NASA- TM-81206). Moffett Field, California: NASA Ames Research Center, June 1980c.

Wilson, W.B. The effect of prolonged non-flying periods on pilot skill in performance of a simulated carrier-landing task (Master's thesis). Monterey, California: Naval Postgraduate School, September 1973.

Woods, D.D. Decomposing automation: apparent simplicity, real complexity. In R. Parasuraman and M. Mouloua (Editors). Automation and human performance: theory and applications. Mahwah, New Jersey: Lawrence Erlbaum Associates, 1996, 3-18.

Woods, D.D. and Sarter, N.B. Learning from automation surprises and "going sour" accidents: Progress on human-centered automation. Columbus, Ohio: The Ohio State University, Institute for Ergonomics, 19 January 1998.

Wright, R.H. Retention of flying skills and refresher training requirements: Effects of nonflying and proficiency flying. Alexandria, Virginia: Human Resources Research Organization, December 1973.

Wylie, C.D., Dick, R.A., and Mackie, R.R. Toward a methodology for man-machine function allocation in the automation of surveillance systems (TR 1722-F). Goleta, California: Human Factors Research, July 1975.

Yeh, Y.C. Safety critical avionics for the 777 primary flight controls system, Space: Aviation's Next Frontier 20th Digital Avionics Systems Conference Proceedings, 2001, 1, 1.C.2-1 - 1.C.2-11.

Yeoman, K.E. Shortcomings of the ATR-72 accident investigation and disposition of the case (A99-16378). 37th AIAA Aerospace Sciences Meeting and Exhibit, 1999.

Young, L.R. An adaptive manual control. IEEE Transactions on Man-Machine Systems, 1969, MMS-10 (4), 292-331.

Youngling, E.W., Sharpe, E.N., Ricketson, B.S., and McGee, D.W. Crew skill retention for space mission up to 200 days (F7666). McDonnell-Douglas Astronautics, Eastern Division, December 1968.

## Multi-stage approach to control room validation

**By Jari Laarni, Leena Norros (VTT)**
**and Leena Salo (Fortum)**

When people are talking about complex technical systems, a nuclear power plant has been typically presented as an ideal example. Controlling this kind of complex system (or system of systems) is a challenging enterprise. Because the final product is complex, it is also a complex task to design a new I&C or CR to a NPP or to upgrade an existing one. To manage this complexity, a more agile, lean and continuous engineering approaches are required. However, there is nothing new under the sun, and even "dummies" know that: according to a recent, tiny book titled Continuous Engineering for Dummies by IBM Ltd, sequential product development is outdated, and more continuous engineering approaches are needed. It is, for example, stated in this book that in order to tame the complexity, continuous verification and validation of the design outputs is needed, i.e., you have to measure twice in order to cut only once.

We have advocating in several papers this kind continuous and phased/multi-staged verification and validation (V&V) of NPP CR systems. It has been proposed that phased V&V is conducted in several stages in sequence so that cumulative evidence of the validity of the new CR systems is achieved. The approach provides continuous support for ongoing iterative design of CR systems by producing regular feedback to design. In addition, by accumulating validation evidence over test activities more thorough evaluation of the CR systems is possible.

The phased approach to V&V is well suited to continuous iterative systems engineering process, in which human-system characteristics are incrementally discovered during the design process; there is a cyclic ongoing specification of system requirements and design solutions; and system requirements and design solutions are concurrently developed.

This approach to V&V has some key characteristics. First, we have emphasized a life-cycle perspective on V&V. Put it simply, there are assessment activities before integration validation tests, and there are also validation test activities after the integrated validation of design solutions. More generally, the life-cycle of a control room extends over decades, and it consists of several developmental stages (see figure below). From a technical point of view, the lifecycle of a NPP CR can be conceived as a progressive evolution from a conventional analog CR through to a hybrid configuration equipped both analog and digital systems and then finally to a fully-digitalized CR. From the perspective of systems engineering, periods of relatively slow evolutionary technical development and bursts of rapid technical change follow each other. During periods of evolutionary technical development, operating experience reviews and performance monitoring activities provide evidence of the validity of the CR systems; in the context of I&C and CR upgrade projects, more definitive validation activities are conducted.

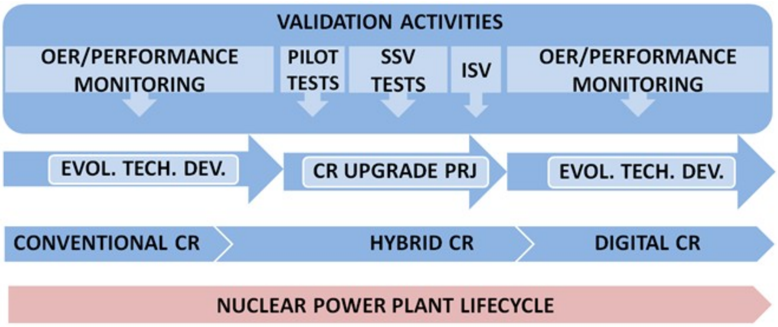
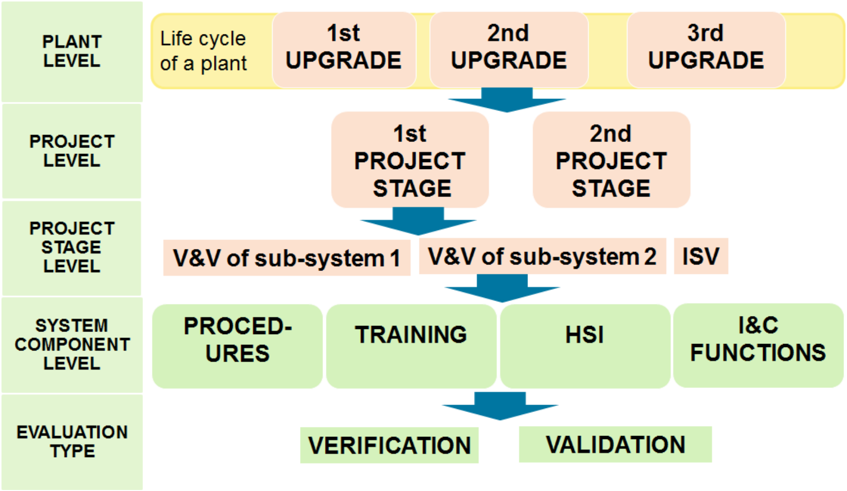


*Figure. Depiction of life-cycle phases of a CR (adapted from Laarni, Salo et al., 2014).*

Second, as the name of the approach suggests, the V&V process is divided into several steps which focus on different parts of the control room as well as on the integrated control room. Phasing can mean different things in this context. In an I&C/CR modernization project, at the highest *plant level*, the term refers to different upgrades of automation and CR systems during the life-cycle of the plant; at the next highest *project level*, it refers to different project stages that are included in a single upgrade (categorized, e.g., in terms of whether the focus is on reactor or turbine side automation); at the *project stage level*, it refers to test sessions that follow each other and in each of which a different set of CR systems are assessed; and at the lower *system-component level*, the term refers to individual elements and components of CR work that can be evaluated in a single validation test session and that can be either verified or validated (*evaluation- type level*). Automation and CR design are often implemented in stepwise fashion so that in a full-scale automation modernization project, reactor and turbine automation are upgraded in successive stages. Within a particular stage of the project, the CR V&V process can consist of several validation steps which focus on different aspects of the CR and operator activity, and each test is focusing on a particular set of elements (i.e., HSIs, procedures and operator activities). In a recent paper we used the following figure to illustrate the differences between these meanings:

**Figure. Hierarchical breakdown of V&V targets (adapted from Laarni, Salo et al., 2014).**

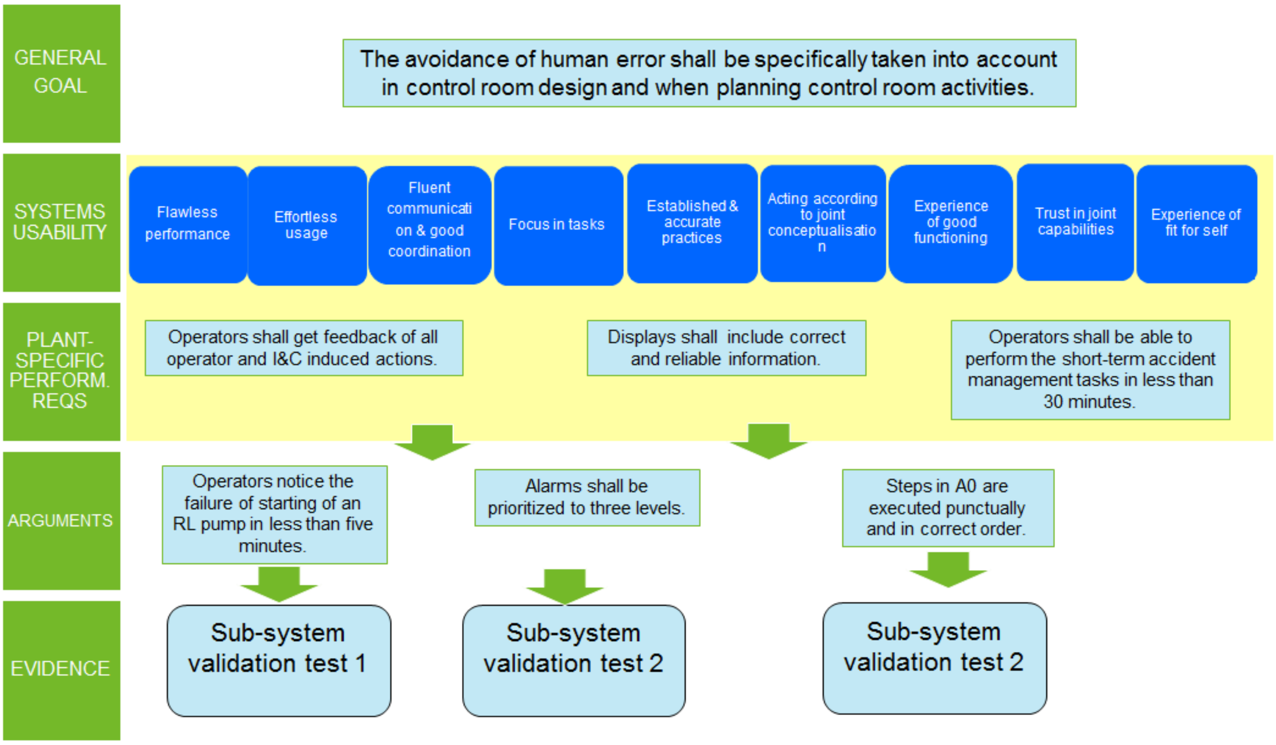


Third, one of the key characteristics of our approach is its comprehensiveness: in addition to design solutions such as HSIs and procedures, also other artefacts such as requirements, style guides and training programs are included in the scope of HFE V&V. As a result, the approach enables a better integration of V&V to design and in other HFE activities and in that way a more complete safety demonstration of the new CR systems.

Fourth, a graded approach is followed in focusing of V&V activities. According to the graded approach, the assessment of HSIs and other CR systems are tailored according to some critical dimensions, such as safety criticality, complexity and novelty. For example, systems and solutions that are safety-critical and that have the highest novelty value will be evaluated most comprehensively.

Fifth, our approach is requirement-based. By this term we refer to a procedure in which control room requirements are systematically used as a reference in the assessment of acceptability of the control room solutions. Verification and validation of CR systems is a requirement-based activity in which a general claim of system safety is further divided into design requirements and theoretically derived requirements. Plant- specific requirements provide one reference for evaluation; more theoretically-derived requirements are based on the Systems Usability concept. Arguments that explicitly argue specific claims about the design are derived from these requirements, and they are then tested in specific validation test activities. In figure below, the goal structuring notation system (see, e.g., Sproggs, 2012) has been used to illustrate the relationships between, requirements, arguments and evidence.

**Figure. Requirement-based approach (Laarni, Salo et al., 2014).**


GENERAL GOAL
The avoidance of human error shall be specifically taken into account in control room design and when planning control room activities.
SYSTEMS USABILITY
Flawless performance
Effortless usage
Fluent communicati on & good coordination
Focus in tasks
Established & accurate practices
Acting according to joint conceptualisatio n
Experience of good functioning
Trust in joint capabilities
Experience of fit for self
PLANT-SPECIFIC PERFORM. REQS
Operators shall get feedback of all operator and I&C induced actions.
Displays shall include correct and reliable information.
Operators shall be able to perform the short-term accident management tasks in less than 30 minutes.
ARGUMENTS
Operators notice the failure of starting of an RL pump in less than five minutes.
Alarms shall be prioritized to three levels.
Steps in A0 are executed punctually and in correct order.
EVIDENCE
Sub-system validation test 1
Sub-system validation test 2
Sub-system validation test 2

Because a lot of data is collected throughout the life-cycle of the plant, systematic methods are needed for the accumulation and systematization of validation evidence and drawing conclusions from the evidence. During the development of the multistage HF validation approach, it was reasoned that a case-based approach is particularly suitable approach when the system to be evaluated is unique and comparison to other systems is difficult. It also supports formative evaluation, in which the interest is to steer the development of the system by successive evaluations, and nevertheless remain independent of the design process itself. Notation systems such as the one shown above can be used in the systematization of requirements, arguments and validation evidence.

Our approach has several strengths over a more traditional approach based on the assessment of CR validity “cross-sectionally” through a single set of ISV tests. The approach provides a lifecycle-based monitoring of CR functionality and usability. The longitudinal process of system validation provides a progressive monitoring of potentiality and usability of the new design. This kind of longitudinal approach to human-system interaction has many strengths over the cross-sectional approach, since longitudinal methods enable to study how the operator performance changes over time in the new CR. Running the same behavioral condition or scenario and measuring particular aspects of operator performance at different stages of system design, it is possible to acquire information of how the potentials of new HSIs have been actualized.

Some experts may now raise their eyebrows and say: “That sounds a pretty good idea, but it does not work in practice. We simply have no time and resources for such a huge effort. One integrated validation test at the end of the design process must be enough for the demonstration of the validation of the design.” What we propose instead is that our approach helps utilities to save money, effort and time, for example because necessary changes can be introduced early enough. – But it is our task to persuade you that this is the correct way to proceed.

## White Paper-WGHOF TWG-
## Dina Notte

To understand the way this white paper has been done, the reader should know:

- I started from "*The recommended topics for participant White Papers*" provided by D. Desaulniers in April 2014 which follows the pre-programmed structure of the workshop.

- I deliberately present my ideas on the subject under a personal point of view rather inspired by my experience.

### Topical Session 1 - Defining the Scope and Objectives of a Control Room Validation

To me the objective of a control room (CR) validation is twofold:

- Validation of the MCR itself encompassing each individual element or part of it
- Validation of the overall socio-technical system as a hole

### Objective of control room validation

The first objective is ensured by several V&V steps including 2D/3D evaluations during detail design and finally under a full scale mock-up of workstation arrangement and MCR lay out. This is done most of the case in wooden mock-up and allows evaluating in "real world" the overall MCR organization, starting from anthropometrics assessment, up to communication, circulation, and escape route appraisal. Reduced scale mock-up can also be realized to evaluate the overall control centre organization if the MCR is embedded in such an architectural option. This approach usually provides good confidence to the final MCR design validation/adequacy according to Human Factor's criteria, although some weak point can be noticed for environmental factors like lighting. Potential rooms where the full scale wooden mock-up will be installed for testing are usually not equipped with the final lighting system and this can be also the case for the FSS. Consequently, typical lighting problems as glare cannot be evaluated properly. Lighting mathematic model permit to overcome this problem but the simulator's fidelity compare to future lighting atmosphere is not yet demonstrated completely. Advanced technologies like virtual reality can provide additional confidence to the compliance of the MCR design to HF criteria. To my opinion this last technique is rather complex to develop and dedicated specialists are needed to master this technique which turns finally to be quite expensive considering its added value in MCR design/evaluation compare to wooden mock-up. On the other hand VR can be an effective method to validate working situation (work place, workstation, accesses, radiation issue, etc) during NPP building design especially for maintenance personnel. Finally I would say that MCR validation should have been done prior to FSS usage and should not be a dedicated objective during ISV, knowing that the MCR at this step of commissioning is already erected and that little degrees of freedom are left to face important discrepancies. I would add that investing time and resources to this objective is counterproductive as test design and data gathering are sufficiently complex when performing an ISV. This way of thinking is also applicable for remote shut down station and remote control rooms on the unit like waste treatment control room or chemical sampling control room. My point of view concerning HSIs and procedure is different because they are both subject to frequent adaptation during the design life cycle and they are flexible by essence. This is true up-to and even after the ISV. The question for HSIs is then to define to which extend we have to concentrate on usability issues which normally should have been detected and solved earlier in the project.

The second objective concerns the evaluation/validation of the overall socio-technical system performance. For NUREG 6393 (ref 1) the general objective of an ISV is *« To comprehensively determine that the design conforms to HFE design principle and that it enables plant personnel to successfully perform their tasks to achieve plant safety and other operational goals ».* This definition presupposes that MCR and the overall design are part of ISV objectives. But when look more precisely in the NUREG 6393, MCR and HSI are considered but mostly in the test bed design as to be conform to criteria like completeness and fidelity but not really as the central subject of this final evaluation. If we consider IEC 1771 (ref 2), the title of the document is: "*NPP- Main Control Room- Verification and Validation of the design*" which express explicitly the MCR design as the centre of evaluation which is not the case of the NUREG 6393 title. So this leaves an open space for discussion.

**Scope of the control room validation**

When running practically an ISV, there could be some gap between the expectations of safety authorities focusing on safety issues and utilities objective concentrating also on plant availability, productivity and industrial unit preservation which raised the scope of maintenance. One should not forget that TMI accident did happen under normal operation. Some other objectives concern the extension of ISV context to remote shut down station, technical support center and remote actuations done by field operators which in deed are difficult to include in ISV sessions, questioning the representativeness of tests conditions and the evaluation of emergency management readiness. This is particularly significant if accidents beyond design are considered like FUKUSHIMA.

Main questions emerging from this first topic concern the way control room can be envisaged as a joint cognitive system rather than a collection of juxtaposed subsystems and people that are assessed independently.

**Topical session 2 - Rationale for Selecting Measures and Acceptance**

**Criteria Selecting Performance Measures and observation**

We could start the discussion from an extreme perspective considering that only one measure could be sufficient to provide confidence, that's to say: the total number of scenario successfully achieved on time. The problems if adopting this position are

- To define a decision method if result is fewer than 100%. To which extend is it acceptable to have a 90% or 80% ratio? One could argue that it depends on scenario severity for safety issue. Another one could make up his mind following the type of weaknesses the operating crew has demonstrated while running the scenario.
- To find what are the reasons why some scenario failed because no other metrics where foreseen and this is particularly true if the problem relies on training.
- To be in line with the systemic Individual, Technology and Organisation (ITO) approach privileged by
  Human Factors like required by the IAEA, considering all the components of a complex socio-technical system.

The only way to solve the question is to consider the acceptable ratio is 100 but some scenario can fail due to external factors (i.e. FSS weaknesses, inadequacy of procedure) and if the FSS is not available anymore, then the scenario cannot been replayed and acceptance criteria is questioned again.

Trying to figure out what could be a possible framework of observation and measures for ISV I have organize this section following 4 steps

- Start from the NUREG position on the subject
- Look to the literature review of Valerie Gawron
- Consider previous review I did on the subject
- Propose a possible Measurement Matrix

Start from the NUREG 6393 position on the subject

Following NUREG 6393, performance of plant personnel should concern following aspects:

- Personnel tasks
- Cognitive workload
- Situation awareness
- Anthropometrics and physiological factors
- Personnel behavior

Literature review of Valerie Gawron (ref 3)

She will be part of the R&D panel of our ISV workshop and proposes in her important literature review the following measures categories:

- Human performance
  - Accuracy based on errors measurement techniques and on percent correct
  - Time execution metrics
  - Tasks batteries
  - and domain specific measures
- Human workload
  - Stand alone Performance measures
  - Secondary task measures
  - Subjective measures
- Situation awareness
  - Performance measures
  - Subjective measures concerning operators and crews

Among the review I identified the following measures that could be considered during ISV:

- Time to complete and Task load. The first *is the duration from operator's first input to the last response*.

  What is interesting to notice is that that when experts are estimating the time it would take to perform

  the task, the correlation between time estimation and real time was highly correlated (+0,98). The second concerns *the time required to perform a task divided by the available time to perform the task*. This method needs to gather objective measurement criteria for identifying the starts and the end of tasks. Studies in aviation reported that takeoff, landing and air refuelling present high task load. This approach could be considered in NPP operation as it is well known that certain plant situations management suppose high task load. In a complementary way, domain specific measures which assess abilities to perform a family of related tasks can be considered in a battery. These measures assume that abilities and effects vary across segments of a mission and are often estimated from parameter describing aircraft states (i.e.; takeoff and climb, cruise, approach and landing, etc) and consecutive frequency of human error. These approaches to efficiency measurement are most of the time applied during ISV.

- In the human performance metrics I found also an interesting approach considering team communication to assess fighter pilot team in military operation (United States Air Force C2 tactical operation). Speech and communication among the crew was encoded under following families: Request, Provide, Strategy, Encouragement, and Fatigue. The codes were applied with 95% agreement among coders, which shows the robustness of the method. In France we call this approach *functional communication analysis* particularly used by R. Samurçay and J.Rogalski. I have used this method several times and particularly in operating theatre during a feasibility study in telemedicine surgery (ref 4). This method is quite powerful to show the learning curve of surgeons and the way they gain ability to pilot and to master the surgical robot. Although I find the method pretty interesting especially in cognitive ergonomics and team performance evaluation, I must admit that it is quite heavy to develop due to communication post coding effort. It is indeed very time consuming and rather luxurious to apply in ISV. Additional I would say that this method is not really applicable to nuclear operation because communications are not strongly regulate by operational rule like in the military field or in the Air Traffic Management sector.

- Among subjective measures of workload, I identified the well-known NASA TLX and the SWAT as multidimensional metrics based upon sub-scales. In that category, the Pilot Subjective Evaluation (PSE), for use in the certification of the Boeing 767 aircraft), is oriented towards work load estimation according to operation (i.e. navigation, FMS operation and monitoring, Engine airplane/system operation and monitoring, manual flight and path control, communications, command decision, collision avoidance). This rating method proved to be useful in identifying design deficiencies. The fact to find metric related to operation rather than on subject experience of workload like the NASA TLX and SWAT is interesting. The former is based directly on tasks as the second is an integration of weighted subjective responses (emotional, cognitive and physical) which in turn, are driven by perception of task demand. Although this approach is interesting, I am wondering if NPP operation is evident to cleave in generic tasks like for aircraft. The more we want to stick as close as possible to operation, the more the rating scale could be complex to fulfil especially in an instantaneous mode during ISV breaks. Hence, the fact that I did not consider *Decision Tree Subjective Workload Measures* which appears to me too complicated to use in FSS environment conditions. That’s why I was also interested by single number rating scale like the *continuous subjective assessment workload*, the *dynamic workload scale* (as tool developed for aircraft certification and used extensively by AIRBUS industries) and the (ISA) *Instantaneous Self Assessment* (used for Air traffic Control by EUROCONTROL). These are all sensible and reliable technique easy to use during breaks and ISA is offering a real advantage: the correlation between operator’s rating of observer’s rating is +0,80 which enable to use subjective technique for both operators and observers with acceptable confidence. That’s why I will consider these techniques later in this section.

- Considering measures of situation awareness, SAGAT is emerging in top of the list to be a reliable and valid tool to evaluate human performance efficiency and effectiveness. SART is also addressed. SAGAT has been adapted into SACRI (Situation Awareness Control Room Inventory) to evaluate nuclear power plan operator’s SA. Studies where launched by HRP. SACRI (Ref 5) is composed by two sets of question on process state parameters: one on primary circuit (19 items) and another on secondary circuit (20 items). Operators have to indicate if those have increase, decrease, drifted or are stable. Conclusion was that SACRI was useful in design assessment when used with time and accuracy measures. As far as I can see, I found this method quite time consuming and cumbersome for operators to fulfil, may be that’s why HRP moved to other measurement like SCORE or PUAT described next.

Personal review prior to prepare an ISV Work Instruction Document

Efficiency

- SCORE (Supervisory COntrol Rating Score (ref 6) is a method used by HRP to evaluate effectiveness.
  The support of the evaluation is two parts: One is dedicated to the scenario summary encompassing: a) Briefing elements, Scenario description elements and validation objectives which can be used to assess a global evaluation; b) Effectiveness measure. SCORE is adapted to scenario specificities and is based upon direct queries (like SAGAT method) and objective effectiveness rating scale on six point's scale which is fulfilled by expert subject matters.
- SAGAT (ref 7). Direct queries. Direct queries techniques allow asking specific questions to operators in order to evaluate their situation awareness of the current state of the plant through specific variables of the process (i.e. T°, P, volume, time needed, system state, anticipated plant state expectation, etc). Queries are covering the three dimensions of situation awareness which are: perception, comprehension and projection which are defined by deep cognitive tasks analysis. The direct queries forms/questionnaires must be fulfilled during scenario tests break when freezing the simulator. Forms analysis provides success scores calculated from a reference matrix prepared by expert subject matter. SAGAT technique has been proven ineffective following potential bias:
  - freezing do not affect performance
  - recognition and memories does not appear as a problem
  - rationalization and generalization tendencies when using SAGAT after the end of scenario is
    not questionable. This hazard is most likely not an issue for this technique as review of
    literature shows those problems occurs when subject are asked to report on how they know
    something, not their assessment of the situation

Subjective rating scale

- Situation Awareness Rating scale (SART ref 8) is based on 14 components data analysis by military pilot
  to be relevant for subjective situation awareness evaluation. Due to the fact that SART items are mixing workload and situation awareness and because SART rating has been founded correlated with operator's performance and workload measures, it does not seem to be appropriated for ISV purpose. Additionally it presents some bias in the evaluation of human performance measurement and as SART result from military objectives which is probably not suitable for nuclear power plan domain. By the way SAGAT and SART are not correlated (ref 9).
- Process Understanding Assessment Technique (PUAT¯ ref 10). This tool used by HRP, is a 3 item rating
  scale based on the 3-item Situation Awareness Rating Technique (SART). From return of experience on PUAT usage, HRP has decided using a simple one-item rating of operators own experience of level of situation awareness as the detailed results of SART type of measures can be very time consuming.

Workload

- SWAT (ref 11) identified 3 dominant factors to evaluate workload: mental effort load (i.e. perceived complexity and task complexity), time load (i.e. Time stress and time needed versus time available) and psychological stress load (i.e. perceive effort and task complexity). The evaluation is done on 3 point self-rating scale for the 3 dimensions.

- NASA TLX (ref 12) results from an important study identifying major factors contributing to cognitive workload. The evaluation is done through 6 dimensions on a 20 points self-rating scale. SWAT and NASA TLX provide similar dimensions to evaluate cognitive workload which prove the credibility of their identification of important factors in subjective workload assessment.
- SWAT and NASA TLX techniques (ref 13) have proven to be unobtrusive and have demonstrated
  sensitivity in variety of multi-tasks environment including Nuclear Power Plants. Additionally, a scientific research whose objective was comparing SWAT and the NASA TLX, shows that the two coefficients were positive and near to one, after Pearson correlations between each performance and each subjective workload measure were calculated to evaluate the concurrent validity of each instrument with task performance The assessment of convergent validity of the instruments showed the high convergent validity of the two instruments considered in this research

Team Work

Teamwork is a broad concept and the teamwork literature (ref 14) points to a number of potential teamwork dimensions that can be applied as foundation for the development of rating scale items. ISV projects should use rating questionnaires for observer and for operator assessment of teamwork that should include:

- Team composition (personality, cognitive ability, motivation, cultural factors, leadership)
- Work structure (team norms, work assignments
- Task characteristics (task type, interdependency, resource management)and especially, multiple and concurrent tasks, uncertainty, changing plans, and high workload that pose difficulties for operating teams in that the team must reconcile conflicting goals and task interference.
- Workload, (attention, time constraints, complex cognitive processes)
- Communication (decision making and consultation, information sharing)

As Salas point it out there remains a need for more robust, reliable, valid, and diagnostic measurement approaches. For example, the pursuit of dynamic and adaptive systems that are sensitive to team-level performance requires unobtrusive and real-time measures of team performance that can be practically implemented, especially in the field.

HRP (Ref 15 & 16) has developed dedicated tool for nuclear industry which name is TEam Self-Assessment. TESA is a crew self-assessment method allowing crews to assess their level of technical and teamwork competence. TESA comprises three parts concerning nuclear operation: One for outage, one for normal operation states and one for emergency which has partly been developed, exposed to user evaluations, and is currently being documented. The preliminary results suggest that even though teamwork competence requirements may be accounted for using global teamwork-competence categories (e.g. building situation awareness), the (sub) competencies and the associated challenge may differ markedly across the three operational states.

Following HRP, the experience with the teamwork results are somewhat similar as with general workload rating scales. The results provide a useful overall indication of teamwork, but general overall rating scales typically capture only large deviations and are difficult relate to specific scenario events. Therefore, the results of overall rating scales need to be interpreted together with the set of measures applied for giving a meaningful input to the evaluation of the control room.

Proposed potential measurement matrix

Finally to rise a possible framework for measurement, I propose to make the difference between objective and subjective measures; and between quantitative and qualitative measures because most of the times these are mix up in test design.

- Quantitative measures refer to evaluation based upon numerical data (i.e. time execution, duration, delay, number of items observed, estimation by expert on a rating scale, etc)
- Qualitative measures refer to verbal or non verbal evaluation (i.e. operator’s opinion, expert’s opinion, communication between the crew, operators behavior being observed by an expert, etc)
- Subjective measures are provides by operator’s self- estimation either via rating scale (and then this is an objective measure), or via explanation during the debriefing (and then this is a qualitative measure). In the present document it is assumed that operator’s self-rating rating scales are considered as subjective
- Objectives measures are provided by concrete data on process evolution or on operator’s task execution and or by expert/observer evaluation. In the present document it is assumed that expert’s evaluations are considered as objective.

So reporting several metrics discussed inhere, we can find there distribution in following figure.

| Measures | Qualitatives | Quantitatives |
|---|---|---|
| Subjectives | 13/14 | 4/6/9/11 |
| Objectives | 12 | 1/2/3/5/78/10 |

Figure: Matrix of Performance Measures

Legend:

1. Human performance efficiency measures like for instance time execution according to time reference and sub-sequence of scenario for instance.
2. Human errors frequency either when piloting the process or when using an inadequate interaction mode on HSIs. These are detected events from process or procedure personnel observation or event that are extracted from the simulator log
3. Human performance effectiveness measure can be assessed in detecting inefficiencies in piloting the plant by expert on a rating scale (i.e. can operators reach a certain point in the scenario, can they keep a certain parameter below a certain value, do or don’t they unnecessarily escalate a situation - e.g. from AOP to Accident, although not “required” per scenario). Effectiveness or inefficiencies can also be used by expert to provide an evaluation on operator’s skills and competence. This measure can preclude the scenario to be removed from the sample making the difference between acceptable performance and deficiencies in training.

4. Human Performance estimation based on operator’s self-rating
5. Cognitive workload evaluated by an expert on a rating scale.
6. Cognitive workload evaluated by operator’s self-assessment
7. Operator’s Situation awareness evaluated by a list of predominant parameter for a given scenario that will be checked by expert
8. Operator’s Situation awareness evaluated from a series of queries upon the process state and important parameters that characterize du current situation. Some additional queries can be added to evaluate the awareness of anticipated event that should occur in the future.
9. Situation awareness evaluated by operator’s self-assessment. These self- assessments can be done with more or less sensitivity according to the evaluation sheet characteristics
10. Team organization, coordination and communication evaluation by expert and operators on a rating scale
11. Usability rating scale done by operators
12. Overall anthropometrics evaluation according to visual observation of operating team behaviors i.e. non-verbal activities which are considered objectively
13. Free comments provided by operators while fulfilling the usability rating scale
14. Debriefing outcomes.

This Matrix could/should also include some evaluation concerning trust and dependency on the operator support system. These notions emerging from advance process system question the trust of operators considering the reliability of sophisticated I&C calculation and the operator’s dependence to the level of automation. These notions are not new since they were pointed out by Bainbridge in her paper “Ironies of Automation” in 1983. The question is rather how to tackle these psychological issues having close relationship with human performance like believes and motivation? In the other hand do we have some confidence in these kinds of metrics considering the specificities of Nuclear Industry and the commitment of personal during training phases before starting a plant?

Although the Matrix Performance measures encompass numerous potential measures for ISV, I would suggest to avoid unnecessary measures and tools and to privilege rating scale notably those fulfilled in parallel by operators and experts on same matters like cognitive workload and situation awareness which can be translated in numerical values.

Without to deny the NASA TLX, which has proven its validity, I prefer to use a simple visual rating scale like discussed previously. The idea is to suppress “noise” around the rating task operators have to perform (i.e. no adjectives, no numerical graduation, and no colours). I have used this ISA looking like method several times in workload evaluation with high correlation among operators and experts estimations.

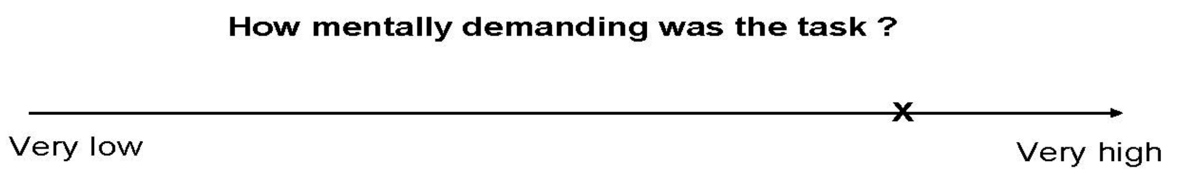


*Figure2: Simple analogical visual rating scale*

Rating is obtained by measuring the distance between the cross and the left side of the scale (i.e. scale should be 10 or 20 cm). I also would draw attention to potential bias in the design of rating scale like the use of uneven scale which favours central tendencies in responses and I am rather surprise to notice how many rating scales evoked in Valerie's book are uneven.

**Performance requirements and acceptance criteria**

Personally, I have no confidence using a benchmark reference to define performance requirements and to provide a rational decision upon design acceptability. Benchmark approach has only a sense when Utility has an important fleet composed by NPP from the same design series (and thus the same MCR design, procedures, HMIs, training, etc) and for which the first of a kind ISV results can be used as a reference for successive ISV performed on the next of a kind. May be Cecilia De la Garza and Jean Paul Barthes (participants to the workshop) can provide some testimony about this very specific industrial scheme.

I have no confidence because we are comparing pears and apples and this is true for new builds and modernization, especially if the MCR is completely redesigned due to advance automation and computerization which is mostly the case nowadays.

To me, it makes no sense to compared human performance from previous design and new design. When evaluating a conventional MCR, time execution is highly influenced by physical workload as a consequence of MCR lay out and conventional panels sizes and design. Old control rooms are often very large and anthropometric is one the most important factor that was optimised/evaluated in the past. On the other hand, information is at disposal permanently and is quickly retrieved in conventional control rooms, which is not the case on computerized workstations. In one case procedure where completely paper based and in new control rooms there are computerized with access and navigation issues. So whether the design is we can gain or loose time anyhow in both design. Conducting a plant in advance control room induces important changes in operator's operating stereotypes and amplifies cognitive demands throughout the entire Rasmussen Model of human information processing steps. An example could be used for vehicles. It seems to me hazardous to use human performance measures based on cars driving reference to evaluate motorbike piloting. Efficiency measured in time execution is not comparable due to adverse effect of traffic congestion on cars for instance.

One must understand that my position concerns the reference for measures, not the reference for scenario selection.

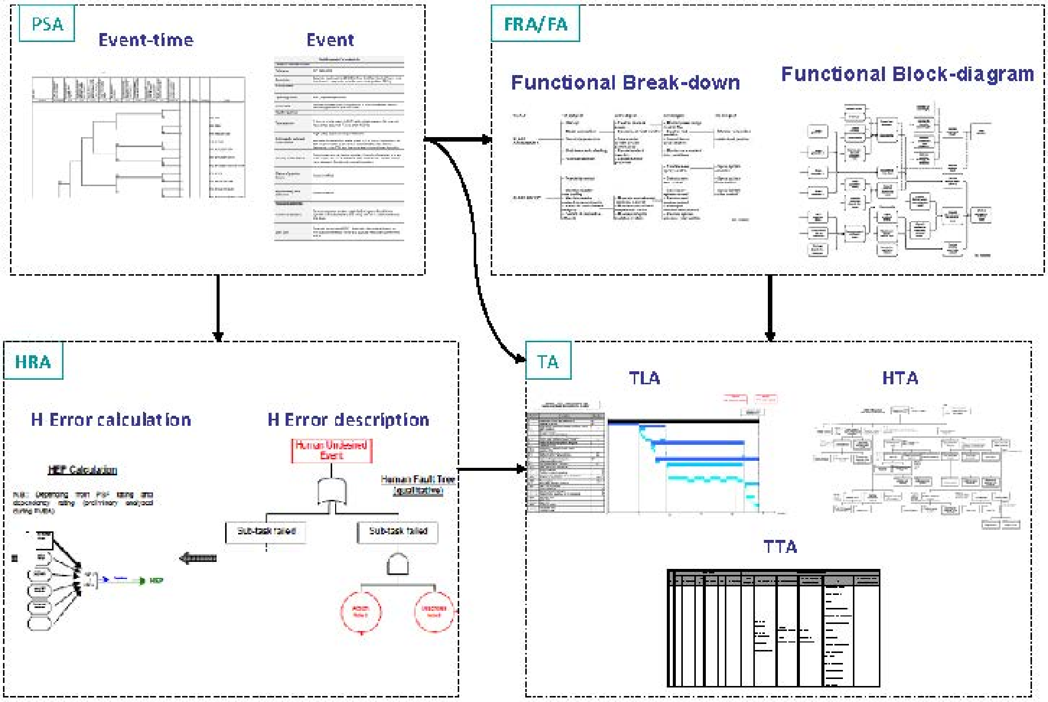
Material from design activities that can be used to inform the performance requirements can be Technical Specification (TS) for normal operating conditions, PSA, HRA, or FRA/FA/TA, for accidental conditions. For both normal and accidental conditions, operating procedure can be used as inputs as well. For the most complex scenario or that encompassing important risk important human action, following inputs are of importance: From this material, efficiency (1 and 2 in previous legend of the Matrix Performance measures) and some effectiveness (3, 7 and 8 in previous legend of the Matrix Performance measures) requirement performance can be estimated and consolidated after having run pre-tests on FSS during ISV preparation.

Main questions emerging from this second topic is a) to define a set of measures that is sufficient in quantity and in quality to prove adequacy and efficiency of the evaluation methodology considering FSS and ISV constraints and available resources and time and b) to define acceptable/proven reference for acceptance criteria.

Establishing the Scope and Content of Test Scenarios

Classical approach in designing NPP is based upon PSA and HRA analysis illustrated at the left part of the figure 3.

**Figure 3: Summary of HFE approaches when scenario selection**



On the other hand NUREG 0711 (ref 17) requires FRA/FA and TA analysis which are illustrated at the right side of the figure 3. By mixing classical and HFE approach:

- PSA outputs provides scenario, accident mitigation timeline and event analysis with consecutive risks evaluation that will be used as basis for HRA, FRA/FA and TA.
- HRA outputs are focusing on human errors and risk significant human action.
- FRA/FA outputs focuses on human-automation sharing
- From these information TA can provide timeline both reactor operator and turbine operator and details on tasks execution and constraints

These complementary methods combined with OER will be useful to support Operation Condition Sampling and scenario definition. The suggested personnel to select scenario as well as the criteria for success or failure are the writers of the output sources, foremost Process & Safety Engineers, Operation and Procedure personnel, Human Factors but also the simulator instructor, and training personnel.

One open question when selecting scenario is to define a sample which is reasonable to run through, during ISV period according to FSS and operating personnel availability. Reduction process in the number of scenario selected can be based upon following criteria and can be challenged:

- Event presenting the most critical risk-important human action and performance shaping factors should be prioritized. Therefore, the scenario selection should take into account the results of the HRA studies.

- Usually a scenario is starting from normal operation following a progressive degradation to incident and accident situations. This permit to reduce the potential numbers of scenarios to build by combination of normal and incidental/accidental scenario.
- Events based on same families of process degradation/operator's actions like LOCA, LOOP, RT, secondary side events, Rods mechanism malfunction, RCV malfunctions, etc. should be merged in one or two representative scenario
- Among limiting accidents, a selection should be made to retain the most severe according to population and environment negative impacts and /or those that have been reported already in international OER
  and for which operators are currently trained on simulator worldwide.
- Event based on multiple field operators' actions could be taken out of the scenario sample, because the FSS environment does not offer sufficient fidelity regarding the real plant operation conditions.
- Scenario should reflect the "real life" of NPP operation. Some events analyzed by PSA studies must be accommodated to present realistic situation which should reduce the sample as well. Operating personnel and training personnel would be able to contribute effectively in doing this filtering.

In case of modernization, there is a strong reference for scenario selection considering previous plant process system and new plant process systems. The ISV will thus focus primarily on what has been changed in the way of conducting the plant. Example, replacing most of the manual commands of the RCV by automation will have an important impact on operator's tasks related to reactivity control. I would say that my primary concerns if performing an ISV of a modernized control room would be the evaluation of the vividness of previous stereotypes and its potential positive (mastering of the process, team routine, pattern recognition, recovering skills due to experience, etc) and negative aspects (control actuation habits in contradiction with new system, loss of situation awareness, usability issue due to computerized interaction, etc). This is particularly true if utility moves from event approach to state approach and again, in this case, reference metric is worthless.

Selecting and Specifying the Test Design

Test design should establish an order of test scenario presentation for each crew and each observer in order to avoid potential problems of bias. For instance a Latin square arrangement of scenario can control for sequence effect but this need to have as many crews as scenario which is impossible to ensure on a NPP ISV. Another approach could consider that all crew should run the same scenario sample. As most of utilities are basing there production on 5 to 6 shift teams and if we consider after reduction criteria a potential set of 30 scenarios that should mean 180 tests to run which is quite a lot. Another point of view could be that teams should run a minimum of two trials series.

But then size reduction is questioned again and additional bias could be encountered like:

- Learning curve effect which can be defined as operator's performance improvement along sequential tests and appearing separately from independent variable effect.
- Fatigue effect (i.e. physical fatigue, decrease of attention capacities, perceptual fatigue, etc) which reversely leads to operator's performance degradation and appears separately from independent variable effect.
- Anchoring effect which concerns variability in observed performance for same task according to the level of complexity of previous task. This effect is due to the scenario sequencing during tests.
- Environmental context (day hour, operator's condition while testing, etc) which can have some influence on operator's performance typically like postprandial effect after lunch especially in warm conditions.

It seems pretty difficult to estimate a preset of reference scenario well balanced with a ready-made test design for an ISV because NPP design projects are quite different in the way concept of operation is developed (including staffing and qualification), how far automation has been spread into the process control, what kind of EOPs are adopted to mitigate accidental situation, which technologies have been chosen for HSIs design and finally what kind of MCR/RSS has been adopted. By the way cultural issues have important effects on human performance, so to me, ISV is a hand sewed product.

**Analyzing the Results and Drawing Conclusions**

I would propose to add also “data gathering” in here because the way data are collected and encoded is important to ensure relevance of results analysis and conclusions drawing. Data gathering is also important to consider according to the number and type of measurements to be done. There is a limit of potential observers because of intrusiveness and thus there is a limit as already said, to the number of observations that can be done simultaneously, in real time. That’s why self-rating (by operators or experts) is the most appropriate because post treatment after ISV session is easy and rapid.

When considering the analysis of results, the central question is to decide whether statistical analyses are mandatory. Coming back the one way metric which could be solely the scenario success ratio, there should be no need for statistical analysis.

The French ergonomic approach of performance evaluation is mostly based upon qualitative assessment under expert’s level of success estimation during debriefing. Cecilia and Jean Paul could give their point of view on this statement. Human Factor’s evaluation as considered in Anglo-Saxon world is rather based upon statistical results in order to provide evidence of results validity like it is respected for psychological laboratory studies. In these experimental conditions, internal and external variables are supposed to be highly controlled which is not the case in real life situations which makes statistical results less robust and more sensible to interpretation. Historically, this schism between laboratory and real world studies has lead to the emergence of the French ergonomic approach centred on activity analysis. The term activity, which as no equivalent in English, can be understood as the course of action. In France if quantitative data are gathered when evaluating the course of action, few of them are object of detailed statistical treatment, with some exception in military, aviation and spatial domains.

Beside these cultural discrepancies, it can be understood that no or few statistical treatment are unacceptable for safety authorities being in charge of license delivery.

Hereunder some remarks to sustain the debate:

- I have noticed very often, notably in scientific congresses, that you may commit different, contradictory or even erroneous conclusions when manipulating data, especially on small sample. That’s why prior to ISV test design, hypothesis should have brought out, in order to be validated or invalided according to results. But this is not enough to guaranty the pureness of data analysis.
- According to the small number of operating crews and scenario sampling, I am wondering if statistics can be applicable under appropriate scientific conditions.
- Applying statistical methods is often considered as THE scientific guaranty of the test/results conclusions QUALITY which in fact ensures most of all stakeholders’ trust and complacency.

Writing this white paper prior to our discussion in Charlotte, I am really wondering what will be the participants' position(s) by end of the workshop, on this particular topic.

**HAPPY TO DISCUSS WHITH ALL OF YOU VERY SOON**

References,

1. NUREG 6393. integrated System validation Methodology and Review Criteria
2. IEC 6171. Nuclear Power Plant- Main control room – Verification ad validation of design
3. W.J. Gawron. Human Performance, Workload and Situational Awareness Measures Handbook. Second edition. CRS Press 2008.
4. D. Notte. Human factor's in Telemedicine. Training surgeons by telemedicine. HCCI San Diego 2009.
5. Development of a situation awareness measure to evaluate advanced alarm systems in nuclear power plant control rooms - D N. Hogg, K. Folles, F STrand-VVolden & B. Torralba. Ergonomic vol 38, issue 11. 1995
6. Braarud O, Eitrheim M. A measurement framework for human factors integrated system validation of NPP control rooms. OECD Halden Reactor Project Report HWR-1063, 2013
7. Endsley, M. R. & Jones, W. M. (1997). Situation awareness, information dominance, and information warfare (No. AL/CF-TR-1997-0156). Wright-Patterson AFB, OH: United States Air Force Armstrong Laboratory
8. Taylor, R.M. Situation Awareness Rating Technique (SART). The development of a Tool for aircrew system design. Situation Awareness in aerospace operation. AGARD proceedings no 478, 1989)
9. Ensley, M. A comparative analysis of SAGAT and SART for evaluations of situation awareness. Proceedings of the Human factors and ergonomic society. 1988
10. Strand, S. and Svengren, H. (in progress). The Process Understanding Assessment Technique (PUAT). Halden, Norway: OECD Halden Reactor Project
11. Reid, G. and Nygren, The subjective workload assessment technique; A scaling procedure for measuring mental workload. (P. Hancock and N, Meshkati eds). North Holland. NY 1988.
12. Hart, S and Staveland, L. Development of NASA-TL: Results of empirical and theoretical research. Human Mental Workload. North Holland. NY 1990.
13. Rubio, S, Díaz, Martín and Puente. Evaluation of Subjective Mental Workload: A Comparison of SWAT, NASA-TLX, and Workload Profile Methods
14. Salas E, Nancy J. Cooke and Michael A. Rosen. On Teams, Teamwork, and Team Performance: Discoveries and Developments of Human Factors: The Journal of the Human Factors and Ergonomics Society 2008 50: 540
15. Skjerve, A.B., Kaarstad, M., Holmgren, L., 2013. "Teamwork competence requirements in nuclear power plant control rooms". In: R.D.J.M. Steenbergen, P.H.A.J.M. van Gelder, S. Miraglia, A.C.W.M. Vrouwenvelder (Eds.), Safety, Reliability and Risk Analysis: Beyond the Horizon. CCR London, UK: Taylor and Francis Group. (ESREL-2013)
16. Broberg, H. (2009). Teamwork in Swedish nuclear power plant operations crews. Master's thesis in Ergonomics Université Paris Descartes
17. NUREG 0711 Rev 3. Human factors Engineering Program Review Model