# Achieving Reasonable Confidence in the Validation of Control Room Designs and Modifications: A Summary of the 2015 NEA Experts' Workshop

**David R. Desaulniers, PhD**
Office of New Reactors
U.S. Nuclear Regulatory Commission
Washington, DC, USA
david.desaulniers@nrc.gov

## ABSTRACT

Defining and achieving reasonable confidence in the validation of nuclear power plant control room designs is a complex challenge for many reasons; not the least of which is that validation is a test of strength rather than a test of truth. In 2015, the Nuclear Energy Agency, Committee for Safety of Nuclear Installations' Working Group on Human and Organizational Factors (WGHOF) hosted an experts' workshop to identify and critically examine potential means to address the challenge of achieving reasonable confidence in control room design validations. The experts' principal task was to answer four challenge questions. (1) What are the critical considerations in defining validation objectives and how do these impact achieving reasonable confidence? (2) 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? (3) 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? (4) How should the validation results be aggregated and analyzed to determine the final validation conclusions? This paper provides an overview of the experts' responses to these challenges and briefly touches upon their implications for future directions.

*Key Words*: Validation, human factors, nuclear, control room

## 1 ACHIEVING REASONABLE CONFIDENCE

Validation is a critical step in the design of nuclear power plant control rooms and major control room modifications. It is important, not just to designers, but to regulators as well. Integrated system validation (ISV) is a method of validation in which a design is evaluated using performance-based tests to determine whether the integrated system design (i.e., hardware, software, environmental, and personnel elements) meets performance requirements and supports safe operation. Although ISV can provide important insights into how well system elements work together to support overall system performance, the complexity of nuclear power plants (NPPs), the diversity of operational conditions, and practical resource constraints collectively challenge our ability to achieve reasonable confidence in the results and conclusions we draw from these tests.

A byproduct of the complexity of NPPs is the large number of conditions and scenarios that can present themselves during the life cycle of a plant and therefore, theoretically, should be tested. At the

same time, the potential for significant adverse impacts on public safety and the environment demands a strict and limited tolerance for failure. Yet tests of the integrated system involving trained crews of nuclear power plant operators and other personnel in full-scope simulations of main control room (MCR) designs are highly resource intensive, with the availability of trained operating crews and simulator time constrained by practical limitations. These conditions challenge our ability to conduct the number of test trials that can support valid statistical analysis of the test data to achieve the desired levels of confidence in the test results; levels typically associated with rigorous analysis of large sample data. These challenges are in addition to questions of how to integrate the results obtained from diverse measures to arrive at a decision concerning the acceptability of the design.

As a result of these and other challenges, achieving reasonable confidence in ISV test results is a matter of particular concern for ISV tests of NPP MCR designs.

## 2 THE APPROACH

In an effort to identify pathways for achieving reasonable in the validation of MCR designs, the Nuclear Energy Agency (NEA), Committee for Safety of Nuclear Installations' (CSNI) Working Group on Human and Organizational Factors (WGHOF) brought together a groups of experts to focus on this important issue. From February 19 through February 21, 2015, the WGHOF hosted an international experts' workshop on the validation of nuclear power plant main control room system designs and modifications. The theme of the workshop was "Establishing Reasonable Confidence in the Human Factors Validation of Main Control Room Systems of Nuclear Power Plants" and it had three high level objectives:

- Critically examine preliminary and final (integrated) validation activities to better understand their strengths, limitations, and potential inter-relationships with respect to the technical and practical considerations for achieving reasonable confidence in nuclear power plant control room designs and modifications.
- Identify recommended practices, potential solutions, and available technical bases for addressing current limitations.
- Identify priority areas for future research.

The workshop participants included 28 individuals invited by the workshop task group based on their recognized expertise in planning or conducting control room validations or validations of complex systems. Their principal task was to answer four challenge questions that the workshop organizers identified as capturing key challenges associated with control room validation. Each question pertained to a major element/phase of conducting a validation. The workshop task group provided these questions to the workshop participants several months in advance of the workshop so that the experts could consider their responses and develop white papers addressing these challenge areas. These white papers were posted to a website open to all workshop participants to facilitate sharing of views and concepts in advance of the meeting. During the workshop, each question was addressed in a topical session comprising challenge presentations followed by small-group brain-storming sessions. The workshop concluded with panel sessions reviewing the key concepts and recommendations for future practices and research in control room validation. The summation that follows recaps the experts' views expressed in response to each of the four challenge questions posed for discussion during the workshop.

# 3 WORKSHOP DISCUSSION HIGHLIGHTS

## 3.1 The Notion of Preliminary Validation

***Beyond ISV*:** During the early preparations for this workshop a consensus emerged among the organizing committee that the scope of the workshop should not be limited to a focus solely on ISV. Rather, review of the known challenges associated with conducting and interpreting ISV tests suggested that a broader examination of the control room design process and methods used to evaluate the design could be fruitful. To communicate this broader scope of interest to the workshop participants, the organizing committee used the terms "preliminary validation" and "final validation." Given that the purpose of including "preliminary validation" in the workshop scope was to explore new paths to achieving reasonable confidence in control room validation, the organizing committee did not define or bound this concept for the participants. Understandably, workshop participants sought to define this term during the workshop and potential definitions and concepts were discussed in each of the topical sessions.

There seemed to be little disagreement among the workshop participants that with appropriate constraints, the results of activities conducted prior to ISV may be useful to achieving reasonable confidence in the validation of control room designs and modifications. However, the participants discussed at length this notion of "preliminary validation" (PV) in an effort to reach a common understanding. The discussions touched upon the definition of PV, the role of PV, its relationship to ISV and life cycle evaluations, and other practical considerations.

***Defining PV:*** The comments highlighted the fact that the notion of PV has yet to be defined and there was no clear consensus on what the nature or role of PV should be. Participants noted that as the term suggests, it is validation that is performed before the final ISV. However, the questioned remained as to whether PV is a pre-ISV or a set of HFE-tests conducted prior to the final ISV. Several of the participants recommended that there be a clear distinction between PV and "every day" design tests.

Participants offered different visions of the role of PV. Some suggested that the purpose of PV is similar to final validation; to demonstrate acceptability, but PV has an additional purpose of evaluating design features. A slightly different but complementary view was that the characteristic feature of PV is that its scope is more limited than that of ISV (i.e., PV does not aim to validate the whole MCR or all of the systems, but only selected parts of the whole (e.g., systems validated separately from each other)). In this conceptualization, PV is a series of test activities distributed over the design process that support periodic validations, throughout the design process.

Much of the discussion about defining PV was concentrated on clarifying the relationship between preliminary validation and final validation. One line of discussion focused on a gap or boundary between PV and final ISV, while another line of discussion focused on an evolutionary process where the PV and ISV stages are defined by the completeness of the design. In the latter view, the completeness of the design determines the characteristics of the validation. At some point in the process of validation activities, a stage is reached where PV stops and ISV begins. This change of stages occurs when the design is complete. But it can be a challenge to define when the design is complete or when it is complete enough (e.g., to support a safety determination). Several considerations were suggested such as: when you can run all of your scenarios; when your procedures are ready; when you have trained crews.

Yet another approach to distinguishing PV from ISV focused on a potential functional relationship between the two. It was suggested that one objective of PV could be to evaluate whether a design is ready for final ISV. Preliminary validation could be a series of independent tests dedicated to validate human-

system interface components independently, one from each other, but considering future ISV activities (e.g., by applying ISV scenarios, real procedures and so on). Additionally it was proposed that PV and ISV may differ in terms of the acceptance criteria that are employed: for preliminary validation, the criteria could be, for example, improvement of performance over time and tests; and for ISV, pass/fail-type of acceptance criteria. Finally, from a broad view perspective of validation, it was suggested that PV should be thought of as a part of a longitudinal and evolutionary process; part of the lifecycle of evaluation.

***Practical Considerations for PV:*** Despite, or perhaps as a result of, the varied notions of what preliminary validation could or should be, it was not a unanimous view that guidance was needed. There seemed to be some controversy as to how much guidance is needed for PV testing. Some thought that more guidance, and some kind of regulatory guidance is needed, some thought that the designers can do whatever they like during the design process. According to this latter view, it was suggested that perhaps designers could benefit from industrial guidance, but that there is no place for regulatory guidance for preliminary validation.

Several comments touched upon team independence as a potential practical challenge to conducting PV activities. Providing an independent validation team can be a resource challenge and establishing an expectation for independence early in the design process would add burden to the validation process. Some participants raised the question of whether expectations for independence of the validation team require rethinking and modification for validations that include PV. For such circumstance it was asked whether a graded approach to independence might be appropriate, and if so, how we might do that.

Considered collectively, the comments and questions concerning the notion of PV suggest that whereas many see PV as a promising avenue towards achieving reasonable confidence, much work may need to be done to better define the concept and its relationship to final validation.

## 3.2 Challenge 1: Critical Consideration in Defining Validation Scope and Objectives

How should the scope and objectives of control room validation be defined in the context of nuclear power plant control room design? As the working group was developing plans for the workshop, it identified this and related questions that it considered important to establishing confidence in the results of control room validations. Ultimately, the working group elected to challenge workshop participants with the following question:

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

To stimulate the workshop participants' responses to this question Robert Hall, REH Technologies, and Julie Reed, Westinghouse Electric Corporation (WEC), gave a joint presentation that cast this question in the context of conducting validations for new plant designs in countries outside the country of origin, thus adding additional considerations for the participants to weigh as they engaged in the subsequent brain-storming.

Mr. Hall and Ms. Reed rejected the conception that there could be a standard plant providing a basis for control room validations in all countries, because regulatory expectations and practices vary from one country to another. However, they see the need for a debate about the possibility of having a universally accepted validation process and optimization of the validation process for cross-country migration. Their presentation raised several topics for consideration, including: (1) what should be considered in setting the validation objectives for designs that may enter the world market; (2) in what ways can validation test

results be used and reused for different purposes; and (3) what are the differences between validation of a "first of its kind" design and validation of a design based on a predecessor/reference design.

***General Considerations:*** In the ensuing workshop discussion of this challenge question, participants identified several general considerations that could affect both the scope and objectives of the validation. These included: (1) the type of project and extent of change, e.g., whether the validation was for a new plant or a modification and whether the changes, in either of these instances, were evolutionary or revolutionary; (2) the operating philosophy/concept of operations, and (3) tests that have been done previously. On this last point it was countered that whereas prior evaluations increase confidence, even if they are not sufficient by themselves, one cannot necessarily take a validation plan from one country to another. Even with the same design, the validation team may need to do specific validations in different countries.

***Validation scope:*** In discussing the scope of control room validation, many participants were aligned with the view that the validation should encompass the overall socio-technical system. While this view had support conceptually, it raised practical considerations. For example, including activities outside the control room that impact the main control room (e.g., maintenance and activities carried out at secondary control areas), was thought to be a good idea that was consistent with defining the scope as the overall socio-technical system, but some questioned how this might be implemented and achieved in practice.

Other recommendations on scope were more concrete such as: (1) establish a check list of most frequent tasks linked with normal operation, difficult tasks with abnormal operation… use this database as a common base to be used by different users, (2) consider lessons learned from past operating experience, and (3) use PRA/HRA to inform PV and include risk-significant actions. Several recommendations focused on the practical (e.g., cost, schedule, and resource availability (people & equipment)). In the theme of practical considerations, some noted that it is important to pay attention to costs and concentrate on important systems and issues in validation tests. However, it was also noted that it is not the goal to make validation cheap and simple if reasonable confidence is compromised and it was questioned whether it should it be simple to demonstrate the safety of a complex system.

It should be noted that the scope of validation was generally not seen as static, though views differed. Some argued that the scope of validation increased as the design developed and others saw it as narrowing. Consistent with this latter view it was proposed that perhaps with a wide range of performance testing in preliminary validation, one would only need to perform limited testing during ISV. This would appear to presume that it is possible to identify tests of sub-systems that would be unaffected by the subsequent integration of the sub-systems and that the ISV tests would focus on areas of performance likely to be impacted by the integration. Related to this notion, it was suggested that to achieve reasonable confidence in conclusions, the focus of ISV should be on key systems and interactions so that sufficient information can be generated concerning those aspects critical to the integration of systems.

As noted under general considerations, it was recommended the operating philosophy or concept of operations be addressed in the scope of the validation. The rationale here is that ISV is not just for a design, but also for a particular operating philosophy/concept of operations. An operating philosophy/concept of operations can be specific to a country, or reflect particular constraints that are required by the operating organization (e.g., level of acceptable automation, crew member versatility and interaction). Consistent with this view, it was recommended that the operating philosophy/concept of operations be tested as early as possible. For example, novel operational concepts (e.g., new staffing concepts) should be tested early on, before it is too late to change them.

***Validation objectives:*** Views concerning the objectives of validation and the critical considerations in determining the objectives were varied. One line of discussion was whether the objectives of design validation can be distinguished from those of design testing. It was proposed that validation is a determination of acceptability whereas design tests explore alternatives/strengths and weaknesses in design. Some individuals suggested that validation is a test of the completeness of a design, noting that it has to be defined when the design can be said to be complete. Many individuals expressed the belief that we cannot determine a fixed set of objectives for validation. In support of this position individuals commented that different types of projects have different kinds of objectives.

Related to the notion that validation is a determination of acceptability, it was noted that to the extent that an objective of validation is attaining reasonable confidence, it is necessary to define what "reasonable" means. In this line of reasoning some proposed that the overall objective should be a human performance centered safety case for which ISV is only one piece of evidence. At a more detailed level it was suggested that one objective of validation should be demonstrating that the testing context generalizes to real world performance. As a cautionary note, it was asserted that clarifying the objectives of validation is important because sometimes there are conflicting motivations (e.g., schedule pressures).

### 3.3 Challenge 2: Performance Requirements, Measurement Selection, and Acceptance Criteria

Approaches to the human factors validation of MCR systems may differ in their requirements or emphasis regarding measurement during test scenarios and what constitutes acceptable performance. However, since most MCR validation projects evaluate quite similar systems, there may be substantial overlap in these areas. Nevertheless, there is currently no universal agreement regarding the measurements, performance requirements, and associated acceptance criteria that should be used for establishing reasonable confidence. In developing its plans for this workshop, the working group identified a number of questions related to this topical area and ultimately elected to pose the following challenge question to the workshop participants.

> ***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****?*

To stimulate responses to this challenge question the brain-storming sessions were preceded by presentations by Dr. Cecilia De La Garza of Électricité De France (EdF) and Per Øivind Braarud from the Halden Reactor Project (HRP).

Dr. De La Garza presented a case study of an EdF control room validation for a new reactor which was carried out as an iterative process that included three preliminary validations and four final validations. Her presentation included descriptions of the objectives and methods for both the preliminary and final validations.

Mr. Braarud's presentation challenged participants to consider what is most important with regard to establishing human performance requirements for an ISV. He asserted that the most important issue regarding human factors evaluation is what should be measured and why. Furthermore, for ISV, he considered the link between human performance requirements and system requirements as central, noting that there must be a clear linkage between "technical" requirements (e.g., reactor vessel pressure, core temperature, readiness of safety trains and safety functions) and human performance requirements (e.g.,

process control actions and plant monitoring, requirements of how work should be performed, performance support tools that should be provided in the MCR).

Mr. Braarud proposed several sources/activities that can help identify and clarify human performance requirements and specify criteria including: task analysis, event analysis, function analysis, training, procedures, HRA, and the operating philosophy/concept of operations.

***Performance Requirements:*** The ensuing workshop discussion of methods and approaches for deriving performance requirements can be characterized as having two general themes or areas of focus; the multi-disciplinary approach and hierarchical functional decomposition (see Figure 1.). It should be noted that these approaches were not discussed as alternatives to each other and in fact could be used in conjunction. A multidisciplinary team (e.g., systems, operations, and HF experts) was proposed for conducting a system functional review and task analysis as a means to define human performance requirements as well as for the analysis of empirical data (both quantitative and qualitative) from observational studies.

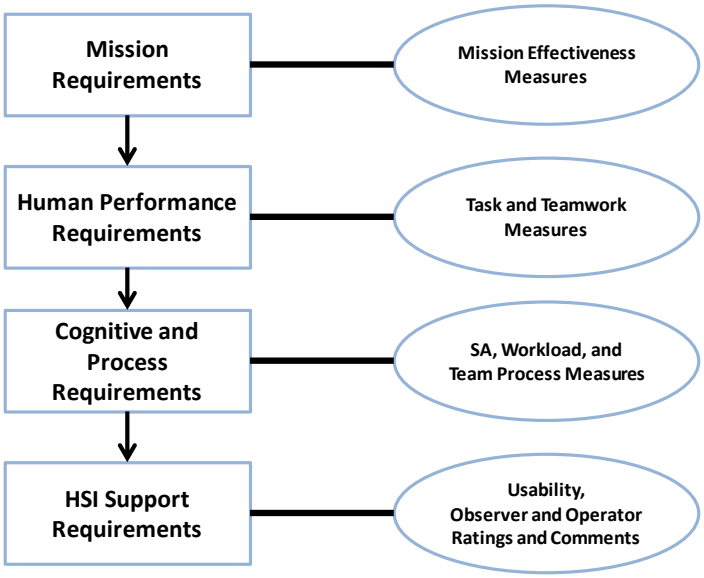


**Figure 1. Hierarchical functional decomposition model relating levels of requirements to types of measures**

The functional decomposition model approach was discussed as having multiple levels of performance requirements including: mission requirements, human performance requirements, cognitive and process requirements, and human-system interface (HSI) support requirements. It was proposed that the criteria at the mission requirements levels should be pass/fail whereas measures/performance relative to lower level requirements should be used as diagnostic indicators to improve design. With respect to measures related to human performance requirements there was discussion but no consensus as to whether they should be pass/fail or diagnostic.

Other potential bases and considerations for establishing performance requirements identified by the workshop participants included: operational experience / previous testing, culture, training requirements, and novel aspects of the design.

***Selecting Measures:*** At a general level, it was commented that the current knowledge base on human performance and measuring needs to be properly incorporated into industry and regulatory guidance for ISV and it was proposed that measures should be tailored to the evaluation phase being conducted as some measures may be more important in preliminary validation than final validation or vice versa.

Whereas there was some debate over the use of practical vs. theoretically-based measures it was noted that it should be possible to have measures that have both practical value and a sound technical basis. Although the participants did not attempt to identify an exhaustive list of the types of measures to consider when conducting validation testing, discussions touched upon the possibility of including: (1) resilience measures (i.e., indicators that should be able to predict good performance in scenarios for which the crews have not been specifically trained); (2) physiological, eye-tracking and neuro-ergonomic measures; (3) usability measures, and (4) observer-based measures. To aid the selection of measures and to improve confidence in measurement results it was suggested that one should use three types of measures to gain convergent validity. Such measures could include: a task performance measure, either human or plant performance, a primary measure of a cognitive or physical construct (e.g., workload), and a secondary measure of another construct that helps provide insight when the task performance measure and the primary measure do not agree (e.g., situation awareness).

***Acceptance Criteria:*** Participants expressed a need for guidance in establishing acceptance criteria. Aligned with this view, there was a perceived need for research to support the development of acceptance criteria in the areas of teamwork, decision-making, command and control, and workload. At a more general level, others saw a need for guidance on when to use different approaches to establishing acceptance criteria, such as determining acceptance based on ability to manage worst case conditions, use of convergent measures, and margins/tolerance. It was suggested that different strategies may be needed to establish acceptance criteria for different dimensions of performance (e.g., worst case approach for workload, tolerance/boundary approach for performance time).

It was also suggested that the operating philosophy/concept of operations for a design could provide a framework to guide the selection of acceptance criteria throughout the design process. It was not discussed how this might be accomplished. Others noted that it may be necessary to tailor acceptance criteria to the phase of validation in which one is conducting the tests.

## 3.4 Challenge 3: Methods for Scenario Development

Additional guidance pertaining to the selection and design of scenarios is a need that is frequently identified for ISV. The questions surrounding scenario development tend to revolve around the circumstances, types, and numbers of scenarios that should be used. This includes guidance for the selection and use of operational conditions, as well as guidance for the identification of what constitutes a sufficient set of test scenarios (i.e., what scenarios should be used and how many of them?). For this workshop, the experts were challenged to answer the following question.

> ***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?***

Prior to the experts brain-storming answers to this question, they had the benefit of presentations from Dr. De La Garza (EdF) and Emilie Roth of Roth Cognitive Engineering whose presentations touched upon potential methods and considerations.

Dr. De La Garza presented an approach that was used by EDF for developing scenarios for the ISV program of a new reactor. The primary focus of the presentation was a three-step methodology that involved an ISV program definition, scenario definition, and participant sampling (e.g., the number of crews to be used). Because the ISV program definition impacts the construction of the validation test scenarios, the first step involves reviewing the principles of an ISV program to ensure that the combinations of different control room components (e.g., procedures, human-machine interfaces, personnel) help achieve the performance objectives of the plant (e.g., safety, production). Dr. De La Garza explained that the evaluation campaigns (e.g., simulations, tests) performed throughout the ISV should cover all of the themes to be validated (e.g., procedures, human-machine interface, team organization) as well as the different possible operating situations, such as normal operations and emergency operations.

She further explained that EdF carries out its validation in an iterative fashion using a series of complimentary tests [1]. An evaluation campaign generally consists of 5 to 10 scenarios, with each scenario typically lasting between 3 to 4 hours during which time the team addresses a variety of operating activities. The principal criteria for selecting the activities used in the campaigns are frequency of use, operating issues, importance, level of innovation, level of complexity, and variety of resources used. She proposed that by working with the regulatory agency and following an iterative process a reasonable level of confidence is achieved in the operating system's performance.

Dr. Roth's presentation challenged participants to construct validation scenarios that are designed to elicit human performance responses similar to those that one would expect in real-world situations. According to Dr. Roth, human performance in real-world work settings is a function of three interacting influences: (1) individual and team cognitive and collaborative factors, (2) situational complexities in the unfolding events, and (3) attributes of the available support artifacts (e.g., displays, procedures). In order to design validation scenarios that are authentically accurate, the scenarios must be sensitive to and representative of the convergence and interplay among the three elements of the cognitive triad. Dr. Roth proposed that in actual accidents involving complex systems (e.g., Fukushima Dai-ichi), it is the confluence and interplay of these three complicating factors that could challenge the operators and lead to performance vulnerabilities. She believes that the challenge for the validation team is identifying operating conditions and designing scenarios that go beyond the routine textbook cases to sample the 'edge cases' – "those demanding decision-making situations that constitute the 'edge' of human performance that could lead to performance degradations."

After touching upon several practical challenges to testing edge cases (e.g., limitations in simulator capabilities and time constraints on scenario durations) she concluded by suggesting approaches for identifying edge cases (e.g., start with the critical human actions, similar to the approach used in HRA, and then use experts to decide what could challenge the actions using methods such as operational experience reviews, lessons learned analysis and cognitive task analysis).

***Scenario Content:*** The workshop participants identified a number of sources that could serve as the basis of input to scenario development along with many practical recommendations for the framework or process of developing scenarios. Although many of the recommendations were consistent with or are included in currently available guidelines, the reader is encouraged to reference the full report for the specific recommendations. Many participants stated that NUREG-0711 Revision 3 [2] is quite adequate with regard to scenario development and should be used as the selection basis for building scenarios as it

provides an already established and widely accepted sampling basis. The guidance document lists a total of 15 attributes or considerations that are grouped across three primary dimensions (i.e., plant conditions, personnel tasks, situational factors), which can serve as the foundation for identifying and selecting the operational conditions for integration into scenarios. However, there was an additional perspective offered by some who believed that the scenarios that are worth testing are the ones that: have an absence of key information, exhibit conflict between automation and operator intent or desire, and cause operators to have to plan and then re-plan.

***Edge Cases and Beyond-Design-Basis Scenarios*:** Dr. Roth's proposal to include edge cases stimulated considerable discussion among the participants. One participant identified the H.B. Robinson fire of 2010 as a potential example of an edge case, noting that it was not anticipated and asserting that if crews can be shown to handle such events it gives confidence that the design is robust. However, a clarification was put forth that unexpected events (e.g., beyond-design-basis (BDB) events) are not necessarily edge cases. For example, there can be edge events from a plant perspective, or from a human response perspective. Thus, it was recommended that for testing purposes, edge cases should be defined in terms of operator challenges (e.g., cognitive/workload/team-collaborative/planning).

Participants also discussed the proposal to include BDB cases in validations. Some participants raised the concern that it is very difficult to ask a vendor to perform validation for BDB cases, as it is by definition boundless (i.e., outside the design box). Moreover, it is not clear what to do with the results; perhaps the design changes would be prohibitively expensive. In addition, it was observed that whereas one could argue that being able to handle such cases implies that the plant will perform well for the design basis, such logic may not always prove correct.

At a more general level it was debated whether scenarios should be representative cases (i.e., realistic) or edge cases similar to a 'stress test' that push the limits of the system? One argument for selecting scenarios that are representative was that ultimately we need to generalize. Another argument was that safety is the ultimate goal; therefore, it is most important for the scenario to be informative, not necessarily representative. It was proposed that validation tests should be designed to gain the most information by creating scenarios with "maximum entropy" (i.e., more uncertainty). This notion was explained by noting that if you expect something to happen (e.g., a crew is successful in a validation test), and it happens, then the event yields little information. However, if something happens that you do not expect, then that event yields greater information. Thus it was suggested that tests involving maximum entropy are useful for preliminary validation, where you want to learn from failure, as opposed to final validation, where you want success.

One point of discussion during this session, though fundamental, should not go without noting. Many of the workshop participants opined that the availability of simulators is a limiting factor and is a major driver for being efficient and effective in scenario selection.

### 3.5 Challenge 4: Analyzing the Results and Drawing Conclusions

There is one point on which there appears to be near universal agreement among individuals that have been involved in the validation of nuclear power plant control room designs or the validation of similar complex systems; these efforts can produce huge quantities and diverse types of data. There is much less consensus regarding how to analyze and draw conclusions from this data, particularly where the decision concerns the acceptability of the design for operation. Accordingly, the working group elected to pose the following challenge question to the workshop participants.

***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?***

To facilitate the ensuing brain-storming sessions on this important and challenging question, Gyrd Skraaning Jr. (HRP) and Robert Fuld (WEC) each provided presentations that outlined their unique perspectives concerning the analysis of validation results.

Dr. Skraaning's presentation challenged participants to consider the presumptions of prescriptive (process-based) validation methods which he identified as including the presumptions that: (1) human performance requirements can be fully understood, pre-defined and specified, (2) correct execution of a prescribed process results in clear and convincing validation evidence, and (3) the procedure generalizes across validation contexts. In contrast, he put forth the hypothesis that prescriptive validation methods fail to provide "…the statistical and logical bases for determining that performance of the integrated system is, and will be acceptable" (NUREG-0711 rev3, p.93).

Dr. Skraaning noted that to judge the outcome of a final control room validation, human performance scores from simulator trials have to be interpreted in light of established acceptance criteria. Thus, some observed performance scores are considered acceptable, while others may be unacceptable. However, he asserted that this is only a first necessary step to judge the acceptability of new control room designs. In the later stages of the data analysis and interpretation process, a myriad of detailed validation results have to be organized, weighted, and judged together from multiple angles to reach a conclusion on whether the control room is, and will remain acceptable. He noted that this process, which could be described as an evidentiary approach, is similar to a trial court or a safety case, where structured arguments are used to evaluate a complex body of evidence in order to reach a clear and definitive conclusion; typically in the absence of formal and prescriptive methods. He in turn suggested that there is no universal formula or predefined psychometric procedure that can help us to reach overall conclusions on the acceptability of new control room designs.

Consistent with this line of reasoning, Dr. Skraaning advocated that one should avoid simple approaches to acceptability judgment where the validation team checks if acceptance criteria are met for a selection of human performance measures. He asserted that such psychometrically oriented methodologies may enhance the interpretability of the observed human performance scores, but is useless during the comprehensive decision making process where detailed validation results are compiled, prioritized, and compared to reach a trustworthy decision on acceptability.

Mr. Fuld began his presentation by highlighting difference between the processes of verification and validation. These differences, which are along multiple dimensions, include the processes' empirical goals (truth vs. strength), the types of reasoning they require (deductive vs. inductive), and whether one can hope to achieve a complete answer (yes and no, respectively). The comparison illustrated not only the differences but also highlighted the challenges of undertaking system validation. However, he noted that validation is necessary because, in part, adequate strength of the system must be demonstrated in order to put systems into service.

Mr. Fuld suggests that we should not be skeptical of validation outcomes if: (1) we prepare sufficiently (e.g., per NUREG-0711), (2) we correct problems found during the process and progress incrementally to a refined state, (3) we accept the simulator modeling of safety parameters, (4) we identify a representative test set to rigorously challenge the system on expected (and other) operating

conditions, and (5) we pass or fail on safety criteria, which are objective, conservative, and independently established and confirmed. Rather, he proposes that one approach validation from a "falsification" perspective in which a successful validation is the null result of passing all safety criteria and a failed validation is the failure of some safety criteria. Passing provides little information and only corroborates the adequacy of the current design (which may still fail someday), whereas failing 'proves' that the current design is not yet adequate. Under this line of reasoning, reasonable confidence in successful validation outcomes comes not from the accumulation of successful outcomes (e.g., through repetitions) but rather from performing appropriate preliminaries (i.e., design development).

Mr. Fuld's presentation also touched upon the use of pass/fail and diagnostic criteria. He advocated pass/fail criteria to be based on safety limits and as few as reasonably possible. Diagnostic criteria should be as diverse as possible. He also proposed that signal detection theory might provide a useful model for the analysis of convergence between results for pass/fail and diagnostic measures.

***Aggregating the Data*:** As general challenges to the aggregation and analysis of validation data, the workshop participants noted that there is variability in both crews and the scenarios and they expressed concern regarding whether it was possible to get to the heart of variability in ISV context. They also observed that analysis of the large amounts of data that can result from an ISV is resource intensive and time-consuming. Regarding the scope of the analysis, some expressed the view that pass/fail measures are the first step and only one piece of the picture, noting that it is necessary to build a case by using all the evidence, including human performance measures, to ensure the integrated system is validated.

There was a common view that qualitative data and measurements of multiple constructs is likely to be used in preliminary validation. As a result, there was a perceived need to develop better methods to support use of multiple and diverse sources and types of data in making acceptability determinations.

The theme of achieving reasonable confidence understandably came through in these discussions. Participants raised and sought to address the question of what is reasonable evidence and sufficient quality of evidence? Some asserted that quality of evidence can be assessed by looking at the consistency and relationships between various types of evidence, noting that there can be evidence from various sources and they should be consistent. Does it meet expectations for convergence or divergence with variation in conditions (e.g., workload)? Are the trends consistent with expectations (e.g., deterioration in performance with increasing workload)?

The participants raised the question of how to achieve statistical relevance (inference) with a low number of crews and particularly noted the challenge of dealing with variability in performance in scenarios? Few solutions were offered, although it was posited that it may be possible to get closer to being able to use inferential statistics by doing more, shorter trials, or by identifying elements common to multiple scenarios that could be measured and aggregated across trials and scenarios.

Although not necessarily focused on achieving confidence through statistical analysis, another line of thinking was the aggregation of data over the course of the design development process (i.e., it was proposed that crediting preliminary validation as part of the ISV should be considered, but it was noted that how to go about doing this remains to be determined). In line with this thinking, some believed that a promising approach would be formalizing preliminary validation to make it a more prescriptive process, such as part of the "evidentiary approach" that would help to make a "safety case."

***Analyzing the Results***: One challenge to analysis that was noted in multiple discussions during the workshop was that operating crews will sometimes take unexpected, though not necessarily incorrect, paths during validation tests. Some asserted that unexpected results cannot be covered by predefined

criteria. As a result, there is a need to include bottom up analysis of collected datasets for the purpose of capturing unexpected issues. It was further suggested that the need for bottom-up analysis was applicable in early validation efforts through ISV.

One approach to data analysis that was advocated during the workshop discussions was a staged evaluation model which would use a graded approach and periodic regulatory reviews. To further explain this approach it suggested that the staged approach could start with regulatory review of the operational philosophy, then periodic review of the design and preliminary validations, and finally the ISV.

As noted previously, the use of inferential statistics in validations was seen as a challenge and differing views were offered concerning the role of inferential statistics and the appropriateness of using descriptive statistics to form inferences about performance. The workshop participants' comments concerning statistics included: (1) most inferential statistics may not apply, (2) the logic behind hypothesis testing does not apply, (3) inferential statistics and descriptive statistics have a role in validation, but primarily in preliminary validation, and (4) statistics can be used to analyze the aggregation of longitudinal data, which in turn can also serve as "evidence" to support arguments in making a "safety case." To address concerns with the use of inferential statistics, some participants suggested that there might be alternative approaches and proposed consideration should be given to equivalence testing, structural equation modeling, and response surface methodology as possible alternatives. It was also suggested that Bayesian inference may be a more appropriate method than traditional inferential statistics.

***Drawing Conclusions*:** The workshop participants' discussions concerning approaches to drawing conclusions from validation results reflected a diversity of views and indicated that collectively the group considered both the use of acceptance criteria and the evidentiary approach as useful and not necessarily mutually exclusive approaches. In discussing the evidentiary approach, the participants noted that the burden would be on the regulator to be a qualified judge of the adequacy of the safety case. Some suggested that such a process could be supported using an independent review by an expert body such as a technical support organization. However, it was also noted that care would be necessary if independent bodies were used as a "certification" process for validations and one must question if they would improve the integrity of the program. The approach could promote independence in validations but also add uncertainty in their integrity and performance.

In considering the falsification perspective that was advocated by Mr. Fuld in his challenge presentation, one participant noted that when such a perspective is considered in the context of the evidentiary approach, it is similar to presumption of innocence - you have to prove failure or guilt. With regard to which criteria would constitute failure, it was proposed that if a crew does not meet a mission objective (as described in the functional decomposition model) it's a validation failure. However, at the second level of the functional decomposition model (i.e., human performance requirements), if there is a "major" human engineering discrepancy, the team could be split on whether it's a validation failure. Some expressed the view that if there is one failure and there's no pattern, it's an aberration, and it doesn't constitute a failure of the validation. In contrast, if there is a fundamental flaw in this design, some pattern of evidence that supports failing the design should exist. Such a view appears to be generally consistent with the multi-path approach to analysis [3] which separates data analysis (e.g., looking for systematic patterns, aggregating data at multiple levels) from acceptability analysis (i.e., the approach discriminates between how we judge observed performance scores and the judgment of the acceptability of the overall system).

## 4 IMPLICATIONS AND FUTURE DIRECTIONS

The format of the workshop focused on the elicitation of expert views in small group brain-storming sessions. As such, it did not lend itself to the experts engaging in in-depth analysis of the specific views and recommendations put forward by their peers. In addition, although some views may have appeared to garner more support or interest in the small group sessions, the workshop format did not involve a formal consensus assessment/building process. As a result, caution must be exercised in extrapolating from the views of individuals or small groups of experts in this workshop to that of the collective. Despite such limitations, the workshop proved to be of tremendous value in generating and compiling innovative thinking with regard to pathways for achieving reasonable confidence in main control room validations. The full workshop report, which is expected to be released by NEA/CSNI in 2017, should serve as valuable resource for regulatory authorities, technical support organizations, plant operating companies, and others, as they consider updates to their policies, guidelines, and research programs. The NEA/CSNI report will document the experts' views and recommendations in more detail than provided here, include and discuss the white papers submitted by the participants (which were not addressed in this paper), and provide an analysis of the workshop results (both the papers and the discussions) in the context of the latest research and practices in control room validation. Accordingly, the interested reader is encouraged to examine the final report once it is available.

With regard to future directions, the workshop has already served as the genesis for a follow-up workshop by the NEA. The workshop, which will be held in June of 2017, will focus more specifically on how conducting validation activities over the course of a design's lifecycle (i.e., multi-stage validation (MSV)) might be implemented to improve validation of an integrated system (e.g., control room). The objectives of this workshop and associated effort are to: (1) better understand and define MSV's critical elements, (2) explore how it can be conducted and documented so as to yield a stronger and/or more efficient approach to validation than currently achieved through integrated system validation testing, (3) identify priority areas for future research. The working group organizing this workshop is developing a white paper that will be the basis for focused discussions among the invited experts.

The results of the 2015 and 2017 workshops are expected to be used as input to at least two important initiatives. The Institute of Electrical and Electronics Engineers (IEEE) is currently developing a new guide, "Human Factors Engineering Guide for the Validation of System Designs and Integrated Systems Operations at Nuclear Facilities," and the working group developing that guide anticipates mining the workshop report for concepts and methods to incorporate in the guide. In addition, the U.S. Nuclear Regulatory Commission (NRC) periodically updates the guidance of NUREG-0711 to reflect advances in the state-of-the-art in human factors. Like the IEEE, staff at the NRC anticipate using the workshop results to identify areas where validation guidance may be updated and where future research to support NRC regulatory mission appears promising.

## 5 ACKNOWLEDGEMENTS AND LIMITATIONS

This paper documents work conducted by the NEA/CSNI Working Group on Human and Organizational Factors. The author is grateful for their sponsorship of the experts' workshop and thus making this paper possible. The following individuals comprised the WGHOF task group which planned, conducted, and documented the workshop: Per Øivind Braarud (HRP); Amy D'Agostino (NRC), Cecilia

De La-Garza (EdF), David Desaulniers (NRC), Stephen Fleger (NRC), Robert Hall (REH Technologies), Jari Laarni (VTT), Dina Notte (ERGODIN for AREVA), Daniel Tasset (Institut de Radioprotection et de Sûreté Nucléaire), and Cyril Rivere (AREVA). In addition, the task group was fortunate to gain the support of John O'Hara (Brookhaven National Laboratory for NRC) and Douglas Hill (Hill Engineering Consultants for NuScale). John O'Hara was the principal author of the post-workshop analysis section of the forthcoming NEA/CSNI report and Douglas Hill served as an external reviewer. This paper draws directly from these individuals' work and they should be credited accordingly. The responsibility for any errors, inaccuracies, or omissions is solely my own. Any views expressed in this paper that are not attributed to workshop participants are my own and do not necessarily represent those of the U.S. NRC or the NEA/CSNI/WGHOF.

## 6 REFERENCES

1. J.P. Labarthe & C. De la Garza, C. "The human factors evaluation program of a control room: the French EPR approach," *Human Factors and Ergonomics in Manufacturing*. **21(4)** pp. 331-349. (2011).

2. J. O'Hara, J. Higgins, S. Fleger, & P. Pieringer, *Human Factors Engineering Program Review Model (NUREG-0711, Rev. 3).* U.S. Nuclear Regulatory Commission, Washington, D.C., USA (2012).

3. G. Skraaning & S. Strand, "Integrated System Validation: The Acceptability Analysis Process," *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. (2015).