BNL-NUREG-62207
CONF-9508177--1

VALIDATING COGNITIVE SUPPORT FOR OPERATORS OF COMPLEX HUMAN-MACHINE SYSTEMS

John O'Hara
Brookhaven National Laboratory
Upton, New York, USA

Jerry Wachtel
U.S. Nuclear Regulatory Commission
Washington, DC, USA

1 INTRODUCTION

Modern nuclear power plants (NPPs) are complex systems whose performance is the result of an intricate interaction of human and system control. A complex system may be defined as one which supports a dynamic process involving a large number of elements that interact in many different ways. Safety is addressed through defense-in-depth design and preplanning; i.e., designers consider the types of failures that are most likely to occur and those of high consequence, and design their solutions in advance. However, complex interactions and their failure modes cannot always be anticipated by the designer and may be unfamiliar to plant personnel [1]. These situations may pose cognitive demands on plant personnel, both individually and as a crew [2,3]. Other factors may contribute to the cognitive challenges of NPP operation as well, including hierarchal processes, dynamic pace, system redundancy and reliability, and conflicting objectives. These factors will be discussed below.

Higher-level functions depend on plant processes, systems, and components. Personnel intervention can occur at different levels in the hierarchy. The interaction of personnel and automatic control system actions may create variability in plant behavior that is not easily understood by personnel. Further, because operators cannot observe the process directly, it must be inferred from a myriad of indicators which provide information about various aspects of performance. Since complex system performance is a property that emerges from the integration of all the components, it may be difficult to predict the performance of the integrated system [4,5].

While actions requiring very rapid response are typically automated, human performance can also be hampered in situations where events move at a slow pace [6]. These characteristics are exacerbated when process disturbances slowly evolve through the occurrence of small human and machine failures, as is typically the case with significant incidents [2].

The overall reliability and redundancy of NPPs can make failures more difficult to detect. When failures actually do occur, operators may not initially believe the validity of the information, instead assuming that alarms or indications stem from other problems such as miscalibrations.

An inappropriate response by operators to conflicting objectives may play a role in NPP accident situations. The demands to maintain power production may conflict with demands to maintain safety, and situations could arise when the trade-offs between these responsibilities are difficult to make. The incident at the Davis Besse plant is one example of this conflict [7}.

To ensure that complex systems support operator performance and that the negative effects of these cognitive challenges are minimized, the integrated system should be tested to identify any deficiencies in design prior to actual operation [3,8]. Although complex systems have historically been "validated" when the reliability and acceptability of their components have been demonstrated, this component level approach is insufficient because it does not address the interaction between components (hardware, software, and personnel). That is, it cannot be assumed that the integrated system will achieve its objectives merely because all of the subsystems and components, in isolation, achieve theirs [9].

An approach to validating the human factors engineering (HFE) aspects complex systems is needed which adequately addresses challenges to plant performance. However, there are few published documents available that provide a discussion of such an evaluation methodology [10]. This need

2.2.1 Methodological Considerations

Validation Testbeds - For complex human-machine systems where failure can be a safety concern, testing actual systems under accident conditions in a real-world environment is not feasible. Thus, the tests have to be conducted using a testbed that is representative of the actual system. The degree to which the model and HSI deviate from the actual design along dimensions that are important to performance determines its representativeness. Representativeness is a function of completeness, physical fidelity, and functional fidelity. Completeness (or scope) refers to the degree to which the testbed represents the entire HSI and not simply selected subsystems. Physical fidelity involves the detailed appearance and layout of features of the HSI such as alarms, displays, controls, and workstations. Functional fidelity reflects the dynamic response and interactive features of the HSI and the degree to which the plant model completely and accurately represents the process. Included in the considerations of the HSI functionality should be the interface management aspects of the design such as display navigation. Considerations for the process model should include the capability to accurately simulate the operating events to be included in the validation tests.

For validation, the main control room (CR) should be represented with high fidelity and completeness. Validation tests may include scenarios where important actions are taken at remote shutdown facilities, local control stations, and other support facilities. It may not be necessary to provide as high-level simulation of these facilities as for the main CR. The decision to represent such facilities should consider the importance of the actions taken at the facility to safety, the complexity of the actions and the HSIs, and the criticality to the overall test of the timing accuracy associated with personnel actions.

Personnel - Integrated systems should be tested with actual users [18,19,8]. However, personnel are a variable aspect of the system and cannot be completely represented; i.e., the entire population of possible operators cannot be included in validation tests. Thus, a sample of participants should be selected that are representative of the plant personnel. To achieve high fidelity representation, the population characteristics, e.g., level of experience, that can be expected to relate to performance variability should be specifically included as sampling dimensions. This will ensure that variation along important dimensions are included in the tests. Selection of participants should also consider the assembly of operating crews, e.g. shift supervisors, reactor operators, shift technical advisor, etc.

Operating Conditions - Operating conditions are composed of plant states/configurations, events that will cause state changes, and situational factors. Since it is not possible to test every condition that is important to the actual operation of the plant, a sampling process is necessary. The selection of operating conditions should provide a comprehensive basis to permit generalization to other conditions or combinations of conditions that were not explicitly addressed by the validation tests. As with participant sampling, the process should ensure that variation along dimensions important to performance is included in validation tests. Several sampling dimensions are described below: plant conditions, personnel tasks, and situational factors. These sampling dimensions are not exhaustive nor are they entirely independent.

<u>Plant Conditions</u> - A range of events that could be encountered during operation of the plant should be identified, including design basis conditions, failure events, transients, and accidents. In addition, beyond-design-basis conditions should be selected to support testing of feasible conditions that may not have been specifically addressed by designers [1,8,9]. While the design basis for NPPs is single failure, experience shows that multiple failures frequently occur. This is addressed through a defense in depth strategy of which operators are a key part. Thus, the test should evaluate the operator's capability to respond effectively to multiple failures (beyond design basis events). These events may be determined from a plant specific probabilistic risk assessment (PRA).

Personnel Tasks - A range of personnel tasks should be identified, including: risk-significant action as defined by task analysis and PRA, interactions within and between main CR personnel and those in outside CR facilities, procedure-guided tasks, and knowledge-based tasks [20]. Knowledge-based activities include those for which personnel must go beyond preplanned responses and use their knowledge of the plant to analyze contradictory evidence, test hypotheses, diagnose failures, plan courses of action, and evaluate consequences of planned actions.

Situational Factors - A range of situational factors should be identified that are known to challenge human performance, such as secondary fault detection, control of difficult systems such as feedwater, automated system override and manual control, workload transition periods, and fatigue periods. Situational factors should also include error-forcing contexts, i.e., situations design to elicit cognitive errors. These factors help assess the system's tolerance to human error and the ability of operators to recover from errors should they occur.

These dimensions reflect characteristics of operating events and not individual test scenarios. The operating conditions should be developed into detailed scenarios which represent combinations of the dimensions described above. It is important that the scenarios have appropriate task fidelity so that realistic task performance can be observed. To ensure reliable use in testing, scenarios should be well defined; e.g., specific start conditions, events and their initiating conditions, communication requirements with remote personnel, and specific criteria for terminating the scenario.

2.2.2 Major Threats to System Representation Validity

System representation validity can be weakened by methodology issues such as:

- Inadequate process/plant model fidelity - inability to accurately simulate important systems and their interactions.

- Inadequate HSI fidelity - incomplete or inaccurate physical and functional characteristics of the HSI.

- Inadequate participant fidelity - use of participants not from the population to which results are to be generalized, e.g., use of engineers or instructors as inappropriate surrogates for plant operators.

- Participant sampling bias - inadequate sampling of the relevant personnel characteristics expected to cause variability in system performance, e.g., use of senior plant operators only. In addition, the following participant characteristics should be avoided: Members of the design organization, participants in prior evaluations, volunteers only, and participants selected for some specific characteristic, such as using crews that are identified as good. These may introduce bias or restrict natural variability.

- Historical population changes - changes in the characteristics of the population to which results are to be generalized which occur after the validation sampling process has taken place, e.g., changes in operator qualification requirements.

- Operating conditions sampling bias - inadequate sampling of operating conditions such that significant demands imposed by operating events are not included in validation tests; e.g., limiting tests to design basis accidents only.

- Inadequate scenario fidelity - failure to represent those aspects of an operating condition that have a significant affect on human performance, e.g., the use of oversimplified scenarios.

2.3 PERFORMANCE REPRESENTATION VALIDITY

Performance representation validity refers to the degree to which validation tests measure performance characteristics important to safety.

2.3.1 Methodological Considerations

Two aspects to performance representation validity are: performance measurement selection and criteria definition. The safety of integrated system performance is multidimensional and many different variables can be selected to measure it. Plant safety is most directly indicated by the plant performance measures (plant functions, systems, and components). While plant measures are essential, they may be insensitive to effects of the design on important aspects of personnel performance [21]. The skill and expertise of highly trained operators can often compensate for inadequate design; however, there may be significant costs to personnel, such as high workload. In addition, plant performance measures may not provide adequate information to determine the cause of inadequate performance.

Therefore, a comprehensive approach to evaluation based in human performance theory is necessary to adequately assess important aspects of human-system performance [16,19,22,23]. Deciding what aspects of the integrated system to measure is a significant consideration. The NPP operator's role is that of a supervisory controller, i.e, plant performance is the result of the interaction of human and automatic control. The operator's impact on plant safety is mediated by a causal chain from the operator's physiological and cognitive processes, to operator task performance, and ultimately to plant performance through the operator's manipulation of the plant's HSI. To adequately perform their tasks personnel need to have a reasonably accurate assessment of the plant conditions and configuration of the HSI. This is based on the cognitive processes involved in developing and maintaining situation awareness. The ability to maintain situation awareness is related to workload (operators typically perform best when the workload level is moderate since low levels lead to boredom and high levels result in performance decrements over time).

Thus, representative performance measures should include: plant measures relevant to safety (e.g., critical safety function margin and technical specification violations); task performance (e.g., task times, errors); situation awareness (e.g., proper assessment of plant states); and workload (e.g., subjective workload ratings). Candidate measures should be evaluated with respect to their measurement properties, such as reliability, sensitivity (data can discriminate within performance ranges of interest), objectivity, unobtrusiveness, and acceptability to participants.

To draw conclusions regarding the acceptability of integrated-system performance, criteria for the performance measures must be established against which the observed performance can be compared. There are several basic approaches to establishing criteria, based upon the type of comparisons that are performed. Integrated system validation will require a combination of these approaches, since the types of performance to be measured are qualitatively different.

Requirement Referenced - Performance of the integrated system is compared with a quantified performance requirement; i.e., requirements for plant, system, and operator performance defined through engineering analysis.

Benchmark Referenced - Performance of the integrated system is compared with a benchmark system, e.g., a current plant, which is predefined as acceptable.

Normative Referenced - Normative referenced comparison is similar to a benchmark reference comparison, however, the performance criterion is not based upon a single comparison system. Norms are established for the performance measure through its use in many system evaluations. The advantage of this approach is that the same measure can be used in the evaluation of different designs.

Expert-Judgement Referenced - Performance of the integrated system is compared with criteria established through the judgement of experts.

2.3.2 Major Threats to Performance Representation Validity

Performance representation validity can be weakened by methodological issues such as:

- Test-level underspecification - inadequate comprehensiveness of the measures such as measuring operator task performance alone.

- Measurement underspecification - inappropriate data collection, i.e., some variables are appropriately quantified by taking several measures at the same time while others are appropriately measured over time. Taking a heart rate measure at one point in time would represent measurement underspecification because it ignores the dynamic changes in heart rate over time.

- Changing measures - changes in performance because the specific measuring instruments or data collection techniques are changed during the tests.

- Poor measurement characteristics - e.g., poor reliability or obtrusiveness.

- Performance criteria underspecified - flaws in the specification of criteria due to inadequate engineering analyses or human performance assessments.

- Measurement-scenario interactions - changes in performance that occur because the measurement technique interacts with the test scenario. For example, questions that are posed to participants to measure situation awareness during a scenario may influence the performance of participants, e.g., by directing them to seek certain information that they would not have otherwise.

2.4 TEST DESIGN VALIDITY

The way in which the tests are conducted can undermine the logical linkage of the integrated system and observed performance. Test design validity is supported when the observation of integrated system performance is accomplished in a manner that avoids or minimizes bias, confounds, and noise (error variance). Shortcomings in test design can (1) alter the relationship between the integrated system and observations of performance, and/or (2) create enough noise in performance data to make results difficult to interpret.

2.4.1 Methodological Considerations

Important aspects of the test design include: the validation team, coupling crews and scenarios, test procedures, test conductor training, and participant training.

Validation Team - A multi-disciplinary team is needed to conduct an integrated system validation. To support objectivity of the evaluation, the members of the validation team should have independence from the personnel responsible for the actual design.

Coupling Crews and Scenarios - The process of determining how the test scenarios are presented to participants involves two steps. First is scenario assignment, determining which crews receive which scenarios. Second is scenario sequencing, determining the order in which each crew receives their scenarios. For example, in most cases there will be more scenarios than participant crews, thus, each crew will receive some but not all scenarios. In research, this is called an incomplete block design [24]. When such a design is used, consideration should be given to balancing the set of scenarios so that each crew receives a representative range and confounding of individual crew performance with scenarios is avoided. This will properly represent the effects of crew variability on important scenario characteristics.

Another type of confounding that can occur is associated with sequence effects; i.e., effects caused by the order in which test scenarios are presented to crews. The order of presentation of scenario types to crews should be carefully balanced to ensure that the same types of scenario are not always being presented in the same linear position, e.g., the easy scenarios are not always presented first.

Test Procedures - Detailed and explicit procedures should be available to guide the test conductors. The procedures should address topics such as how to brief the participants, when to start and stop scenarios, when and how to interact with participants during scenarios, and when and how to collect measures. Where possible the use of a double-blind procedure should be used to minimize the opportunity of tester expectancy bias or participant response bias (see the discussion of threats to test design validity below).

Test Conductor Training - Test conductors should be trained in the use and importance of test procedures. Training should address the potential for bias and the types of errors that may be introduced into test data through the failure to accurately follow test procedures.

Participant Training - Participant training is an essential part of validation and, within the scope of validation, should be very similar to that which plant personnel will receive.

2.4.2 Major Threats to Test Design Validity

Test design validity can be weakened by methodological issues such as:

- Test procedure underspecification bias - failure to provide test conductors with clear and specific instructions.

- Tester expectancy bias - when the collection of data is systematically influenced by the expectations of the testers, e.g., through unintended cues to participants.

- Participant response bias - e.g., participants may want to provide data that they think the test conductors expect.

- Test environment bias - limitations in the test environment to create a realistic operational environment.

- Changes in participants over time - changes in participants over the course of the validation testing due to effects such as learning more about the HSI and becoming more familiar with the testing environment. Inadequate training is one cause of this type of problem.

- Participant assignment bias - systematic bias in the assignment of test scenarios to participants.

- Sequence effects - bias caused by the sequence in which scenarios are presented.

2.5 STATISTICAL CONCLUSION VALIDITY

Statistical conclusion validity addresses (1) the relationship between performance data and established performance criteria, and (2) the interpretation of that relationship.

2.5.1 Methodological Considerations

Logically, the validation null hypothesis is that performance is unacceptable; therefore, the burden of proof is to establish that the design is acceptable. Since performance of a complex task will vary, it is necessary to consider the possibility that observed performance was due to chance and that a different result would be obtained if the tests were repeated. In research this would be addressed through the use of inferential statistical techniques. While these techniques have been recommended for system evaluation [18], several factors combine to make a rigorous statistical analysis difficult to perform for validation. First, because of the need to test the integrated system under a wide range of operating conditions, there may not be sufficient data under constant conditions to provide reliable estimates of population performance parameters. Second, one may not be able to think in terms of deviations from an optimal or mean performance due to differences in operator strategies to maintain acceptable plant performance. As a result, some statistics may be misleading because the individual crews deviate from statistical parameters for different, yet acceptable, reasons. Therefore, validation data may be analyzed through a combination of quantitative and qualitative methods.

Where possible, inferential statistics should be calculated to determine whether observed performance is reliably within acceptable performance envelopes because they enable the quantification of decision errors. Where the statistical assumptions cannot justify the use of statistical tests or where the sample size for a desired comparison is too small, qualitative comparisons of the observed variability in performance and the performance criteria should be made to determine whether sufficient margin exists to permit prediction of successful performance in the actual system. The basis for the determination should be clearly documented.

Another aspect of statistical conclusion validity is the overall sensitivity of the evaluation for detecting that performance would fall within acceptable limits in the real world. Sensitivity can be impacted by low sample size and noise in the data, both of which increase the predicted range of performance. Statistically, this is a problem of low power. Test sensitivity is another reason to maintain a null hypothesis that performance is unacceptable. If the null hypothesis was that performance was acceptable, low power and test insensitivity would work in favor of validating the design.

The degree of convergence of the multiple measures of performance should be evaluated. When all the measures of performance are considered, there should be consistency in the statistical conclusions. Where performance fails the criterion, consideration should be given as to whether it represents a design deficiency, an artifact of the testing process, or inadequate sample size. If the unacceptable performance is due to a design deficiency, consideration should be given to its root cause. To help analyze human performance problems, the operating condition dimensions that were combined to develop the scenarios should be evaluated. If the sampling process had successfully identified the plant and operational characteristics that contribute to the variability of system performance, examining the specific dimensions that make up problematic scenarios should contribute to the identification and resolution of the root cause of performance problems.

2.5.2 Major Threats to Statistical Conclusion Validity

Statistical conclusion validity can be weakened by methodological issues such as:

- Sources of noise - any aspect of the validation tests that leads to increased noise in test data such as inconsistent application of test procedures and measurement unreliability.

- Low sample size - low sample sizes make it difficult to examine the effects of human variability.

3 VALIDATION AND GENERALIZATION

When the following four conditions are satisfied, a basis for inferring actual system performance is established and the design may be considered validated. First, system representation validity is logically supported such that it may be concluded that the integrated system is representative of the actual system in all aspects that are important to performance. Second, performance representation validity is logically supported such that the measures of integrated system performance and their associated criteria reflect good measurement practices and are concluded to be representative of important aspects of performance. Third, test design validity is logically supported such that a comprehensive testing program was conducted by an independent, multidisciplinary team and there are no plausible biasing or confounding effects to make the predictions of system performance ambiguous. Fourth, statistical conclusion validity is logically supported and based upon a convergence of the multiple measures such that it can be concluded that the performance of actual system will be acceptable.

There are limitations to the generalization process. It is possible that not all potentially important factors were identified or that important interactions between system components were overlooked. There may also be implementation differences in the construction of the actual plant, the training of operators, etc. which make the implemented system different from the integrated system that was validated. In addition, integrated system validation will not typically include considerations or influences of organizational factors, such as safety culture and administrative procedure philosophy, which are important to the safe operation of the plant. Therefore, the prediction of actual system performance and the decisions made as to the acceptability of the final design are probabilistic; i.e., with a degree of error.

ACKNOWLEDGEMENT

Views presented in this paper are solely the authors and do not necessarily represent those of the U.S. NRC.

REFERENCES

[1] Vincente, K. (1992). "Multilevel interfaces for power plant control rooms I: An integrative review," *Nuclear Safety*, *33*, 381-397.

[2] Woods, D., Johannesen, L., Cook, R., and Sarter, N. (1994). *Behind human error: Cognitive systems, computers, and hindsight* (CSERIAC SOAR 94-01). Wright Patterson Air Force Base, Ohio: Crew Systems Ergonomics Information Analysis Center.

[3] Reason, J. (1990). *Human error*. New York: Cambridge University Press.

[4] Rosness, R. (1993). Limits to analysis and verification. In J. Wise, D. Hopkin, and P. Stager (Eds.) *Verification and validation of complex systems: Human factors issues* (NATO ASI Series F, Vol. 110). Berlin: Springer-Verlag.

[5] Wieringa, P. and Stassen, H. (1993). Assessment of complexity. In J. Wise, D. Hopkin and P. Stager (Eds.) *Verification and validation of complex systems: Human factors issues* (NATO ASI Series F, Vol. 110). Berlin: Springer-Verlag.

[6] Wickens, C. (1986). The effects of control dynamics on performance. In K. Boff, L. Kaufman, and J. Thomas (Eds.) *Handbook of perception and human performance: Volume II - Cognitive processes and performance*. New York: John Wiley and Sons.

[7] U.S. Nuclear Regulatory Commission (1985). *Loss of main and auxiliary feedwater event at the Davis-Besse Plant on June 9, 1985* (NUREG-1154). Washington, D.C.: U.S. Nuclear Regulatory Commission.

[8] Woods, D. and Sarter, N. (1993). Evaluating the impact of new technology on human-machine cooperation. In J. Wise, D. Hopkin and P. Stager (Eds.) *Verification and validation of complex systems: Human factors issues* (NATO ASI Series F, Vol. 110). Berlin: Springer-Verlag.

[9] Rasmussen, J. (1988). *Coping safely with complex systems*. Paper presented to the American Association for the Advancement of Science, Boston, MA.

[10] Wise, J., Hopkin, D., Stager, P. and Harwood, K. (1994). Human factors certification of systems. In *Proceedings of the Human Factors Society 38th Annual Meeting*. Santa Monica, CA: Human Factors Society.

[11] Wise, J., Hopkin, D., and Stager, P. (1993). *Verification and validation of complex systems: Human factors issues* (NATO ASI Series F, Vol. 110). Berlin: Springer-Verlag.

[12] International Electrotechnical Commission (in preparation). *Verification and validation of control room design of nuclear power plants* (IEC-Draft Standard -5th). Geneva, Switzerland: Bureau Central de la Commission Electrotechnique Internationale.

[13] O'Hara, J., Higgins, J., Stubler, W., Goodman, C., Eckenrode, R., Bongarra, J., and Galletti, G. (1994). *Human factors engineering program review model* (NUREG-0711). Washington, D.C.: U.S. Nuclear Regulatory Commission.

[14] O'Hara, J., Brown, W., Stubler, W., Wachtel, J., and Persensky, J. (1995). *Human-system interface design review guideline* (Draft NUREG-0700, Rev. 1). Washington, D.C.: U.S. Nuclear Regulatory Commission.

[15] Popper, K. (1959). *The logic of scientific discovery.* New York: Basic Books.

[16] Kantowitz, B.H. (1992). Selecting measures for human factors research. *Human Factors*, 34, 387-398.

[17] Cook, T. and Campbell, D. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Boston, MA: Houghton Mifflin, Co.

[18] American National Standards Institute (1993). *Guide to human performance measurements* (ANSI/AIAA G-035-1992). Washington, DC: America Institute of Aeronautics and Astronautics.

[19] Meister, D. (1986). *Human factors in testing and evaluation.* Amsterdam: Elsevier.

[20] Rasmussen, J. (1986). *Information processing and human-machine interaction.* New York: North Holland.

[21] Hart, S. and Wickens, C. (1990). Workload assessment and prediction. In H. Booher (Ed.) *Manprint: An approach to systems integration.* New York: Van Nostrand Reinhold.

[22] Kantowitz, B.H. (1990). Can cognitive theory guide human factors measurement? In *Proceedings of the Human Factors Society 34th Annual Meeting.* Santa Monica: Human Factors Society.

[23] Bittner, A.C. (1992). Robust testing and evaluation of systems: Framework, approaches, and illustrative tools. *Human Factors*, 34, 477-484.

[24] Kirk, R. (1982). *Experimental design* (second edition). Belmont, CA: Brooks/Cole Publishing Company.

DISCLAIMER

This report was prepared as an account of work sponsored by 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. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.