A SUMMARY COMPARISON OF DESIGN EVALUATION TECHNIQUES

Zachary Spielman and Rachael Hill
Idaho National Laboratory
P.O. Box 1625, MS 3112 Idaho Falls, ID 83415
zachary.spielman@inl.gov; rachael.hill@inl.gov

ABSTRACT

Many nuclear power plants in the United States have recently been granted a 20 year extension to their operation license and are considering modernization. The Idaho National Laboratory is committed to successful modernizations and has been identifying and developing candidate technologies for power plants using methods that match the requirements of an integrated system validation. To do so this paper discusses current and potential workload, situation awareness and other performance methods used in evaluating control room design against the criteria defined by the Nuclear Regulatory Commission. The analysis provided is narrowed to the scope of methods surrounding the Light Water Reactor Sustainability research at the Idaho National Laboratory and therefore evaluates only a handful of the numerous methods used in control room modernizing efforts. As such, it was determined the most popular methods for measuring situation awareness, workload, and performance are measurements that rarely measure their intended construct directly and must be inferred. Physiological measures of workload have matured and become more accessible, sensitive, unobtrusive, valid, and objective but are reliant on variable technology. Situation Awareness measures possess little construct validity but eye-tracking technology is a measure progressing in sensitivity, objectivity, and continues to improve in unobtrusiveness. All things considered, this paper has identified a significant need for a more extensive review of the various performance measures used in modernizing control rooms and integrated system validation.

Key Words: Integrated System Validation, Performance Measurements, Situation Awareness, Workload, Modernization

1 INTRODUCTION

This research is a part of the United States (U.S.) Department of Energy-sponsored Light Water Reactor Sustainability (LWRS) Program conducted at Idaho National Laboratory (INL). The LWRS program is performed in close collaboration with industry research and development programs, and provides the technical foundations for licensing and managing the long-term, safe, and economical operation of current nuclear power plants (NPPs). One of the primary missions of the LWRS Program is to help the U.S. nuclear industry adopt new technologies and engineering solutions that facilitate the continued safe operation of the NPPs and extension of the current operation licenses. Currently, of the 99 U.S. nuclear power plants, 81 have renewed their license for another 20 years [1]. Extending plant life however is accompanied by other challenges to keep the plant functional. The majority of nuclear control rooms have equipment built with 1960's technology. The aged equipment is difficult to replace as vendors have discontinued the production of replacement parts and the cost to special order parts can be prohibitive [2]. One solution is to modernize the control room. Modernizing a control room involves replacing current systems with modern technology. Doing so addresses challenges with obsolescent parts and hopefully offers the efficiency benefits associated with modern technology. Improved technology may also increase overall plant and operator performance [2]. An added bonus to updated control rooms is improving operator recruitment. Currently, there is difficulty recruiting new operators with experience

in older analog technology to meet current job demands within the nuclear fleet [2]. Improving plant efficiency, operator performance, and generating more interest to be a reactor operator will contribute to the ability of nuclear power to remain a cost-effective, clean energy source [1].

Two general approaches to modernization exist. One is to update the entire control room in a single effort, but the most common is updating the control room a single system at a time during scheduled refueling outages. Previous research under the LWRS program has identified ways to support system-by-system upgrades [3]. Research has also investigated how new technologies can be effectively integrated into control rooms to enhance operator performance using both common and custom design methods [4]. Such methods facilitate hybrid control rooms, the early stages of the transition to a fully modernized control rooms but are not informative enough to evaluate a design in later stages of a modernization effort. The decision to accept or reject a possible update is made following an integrated system validation study (ISV). The ISV provides a thorough understanding of the affect the upgraded control room technologies have on human performance. Understanding how technology affects human performance in nuclear power plants is complex, and often requires a suite of human performance measures to capture the nuanced effect of technology. The Nuclear Regulatory Commission (NRC) has characterized the qualities a valued performance measure if used to validate a control room technology. The NRC provides guidance for measuring performance constructs such as plant performance, operator performance, situation awareness (SA), and workload [5]. Since the NRC uses NUREG-0711 [6] as a tool to guide reviews, having an awareness of the criteria within is another step toward successfully licensing a modernized control room. However, the benefit goes beyond licensing to ensure that evaluations are predictive of actual plant operations after new technology is implemented. This paper intends to support the LWRS mission by identifying areas of development that improve the chances of successful control room modernizations. More specifically, this report evaluates the performance measures that are currently or potentially being used in LWRS research in relation to the NRC criteria for an integrated system validation. The goal of this paper is to identify characteristic weaknesses in measures used for certain constructs. The weakest areas will help direct efforts for improving or creating new measures.

2 HSI DESIGN DEVELOPMENT METHODS

The strategy for carrying out and documenting a control room modernization is described in a Human Factors Engineering (HFE) program plan. Although its applications do extend beyond modernizations the NRC uses the “Human Factors Engineering Program Review Model” [6] as a tool when evaluating an HFE program plan used in a plant modernization. It therefore can also be used as a reference document when developing a program plan. For example, the document has some detailed guidelines regarding the methods used to validate a system design. Using NUREG-0711 [6] researchers can tailor design studies using measures following the guidelines outlined in the review document. The benefit is helping ensure the outcome of an integrated validation study is more technically sound. Also, incorporating more informative performance measures during evaluation may improve the likelihood of success for the modernization effort.

The goal of the extensive design cycle is to create a design with the best chances of successfully completing an integrated system validation. As stated, many design methods are used to support rapid iteration and knowledge gathering. Most of the following methods are tools human factors practitioners use to guide iterative design to improve system performance.

2.1 HFE Principles

The first step is referencing a range of basic human factors guidelines to develop displays and interaction methods. These guidelines should be the first considerations of a design and range in detail from Neilson’s list of 10 [7] usability fundamentals to NUREG-0700 “Human-System Interface Design Review Guidelines” [8] which are more specific to nuclear power plant control rooms. Not all the

guidelines will be applicable to every design interface. In fact, evidence from evaluation methods may indicate otherwise and require multiple considerations such as workshops, micro-simulations, and full-scale simulation studies.

2.2 Workshops

Workshops are generally a larger scale collaborative effort between human factors practitioners, target users, and domain experts. During a workshop, a design will be explored by all parties, input will be immediately incorporated in the design and quickly redistributed for exploration and interaction. The idea is to produce rapid iterative prototyping using expert knowledge to provide design feedback. Some researchers have laid out step-by-step methods to carry out workshops in ways to gain more information such as the GONUKE method [3].

2.3 Micro-Simulations

Micro-simulations and micro-tasks are controlled tests wherein a small portion of operations are separated and tested using simplified design schemes. Such tests lend answers to specific questions and support quick turnaround decisions in the iterative design process.

Micro-simulations are useful in testing either general operating schemes as broad as the role of automation to basic function interaction (e.g. turning a valve on or off or noticing an alarm signal). Micro-tasks are designed to answer basic questions using quick and simple methods to collect many data points in a short time period [9].

2.4 Full-Scale Simulation Studies

Only full-scale simulation studies are capable of assessing operator performance to the extent described in NUREG-0711 [6]. A full-scale simulation has the benefit of realism. For example, a team of operators are placed in a realistic plant setting and asked to behave as expected during a real plant event. The simulator then functions as a true plant but only within the bounds of the given scenario. That is, minor deviations from expected procedure can be accommodated but serious departure will likely end in interrupting the flow of the scenario requiring a reset to a planned location.

Due to the level of effort and resources required to run a full-scale simulation study it is recommended that only mature designs well vetted by other iterative methods be the subject of scrutiny. Scenarios usually require a minimum of one hour but can continue long after. Also, crews must be composed of operators with real world experience. Such participants can be difficult to track down. When they are recruited, their time is often limited as well as expensive. For a larger break down in the challenges of carrying out a full scale simulation refer to Reference 10. However, it is a full scale simulation that most closely resembles an integrated system validation.

There are still challenges to carrying out a full-scale simulation but the method holds potential as a tool for researchers to begin validating a new system prior to an ISV. Doing so will help researchers identify any deficiencies before large expense and time is invested in the final integrated system validation (ISV). Knowing the deficiencies while still in the design phase will improve the capability of the INL and LWRS pathway to support a successful plant modernization effort. To support the capability of a full scale simulation as an effective test the common and potential measures have been assessed to determine how well the measure characteristics defined by NUREG-0711 [6] are met.

3 INTEGRATED SYSTEM VALIDATION

The purpose of the ISV is to use performance-based tests to validate that “the integrated system design supports the safe operation of the plant” [6]. Three of the main validations are listed below,

however note that many other considerations are taken and not all are even accounted for in NUREG-0711:

1. The HSI is capable of alerting, informing, controlling, and providing feedback to operators
2. Specific personnel tasks can be accomplished within the time and performance criteria with effective situational awareness and workload
3. HSI minimizes personnel error, assures error detection, error recovery capability

Determining if a system meets these criteria requires a method incorporating a suite of measures that complement each other. As a reference NUREG-0711 [6] has broadly defined five characteristics of an ideal performance measure. This paper aims to summarize common performance measures based on those characteristics. The goal is to identify measures that are possibly due greater consideration when developing a suite of measurements as well as some author perceptions regarding commonly used measures.

3.1 Performance Measure Characteristics

Part of performing the ISV is reporting on the methods used and defending the basis choosing each measure. NUREG-0711 [6] explicitly lists five performance measure characteristics that should be accounted for when considering ISV measures. Table 1 is pulled directly from the document and accounts for the performance measure characteristics.

Table I. NUREG-0711 Performance Measure Characteristics

Characteristic	Meaning
Construct Validity	A measure should represent accurately the aspect of performance it is intended to measure.
Reliability	A measure should be repeatable; i.e., same behavior measured in exactly the same way under identical circumstances should yield the same results
Sensitivity	A measure's range (scale) and its frequency (how often data are collected) should be appropriate to that aspect of performance being assessed.
Unobtrusiveness	A measure should minimally alter the psychological or physical processes that are being investigated.
Objectivity	A measure should be based on easily observed phenomena.

An ideal measure possesses each of these characteristics but realistically there are often trade-offs. A measure may lack sensitivity to remain unobtrusive. A construct as difficult to directly observe as situation awareness (SA) requires inferring SA by measuring related constructs. A measure requiring inferences based off measuring related constructs is viewed in this evaluation as having less construct validity. Objective measures are especially difficult when working with humans and studying cognitive phenomena due to the myriad of individual factors influencing operators including; personal stress, amount of rest, diet, or personality differences. Therefore measures that can control for or circumvent such influences are seen as having greater objectivity. Note that the following methods should be investigated experimentally in comparison to one another to best understand how best to combine them and illicit more precise and accurate information.

Common and proposed performance measures used for ISV have been compared to the performance characteristics described in Table 1. Raising awareness of various measure characteristics improves research development and reporting and, thus, overall research quality. The end goal is having the ability to select the right methodologies to measure human performance for different phases of design and validation in control room modernization.

4 CURRENT METHODS FOR ISV

4.1 Basic Performance Measures

Basic performance measures such as timing, accuracy, and frequency have become ubiquitous and are cheap, efficient and accurate to implement in evaluations as technology progresses. There are multiple methods to collecting such information ranging from observers to event logs. These are measures that can be easily automated. The more they are automated the more accurate and reliable they become.

The three basic performance measures called out in NUREG-0711 [6] are strong across all five measure characteristics. Their weakest aspect lies in individual differences and defining thresholds for measurement. Researchers must be clear when defining set points such as stop/start times, when an action has concluded, and accuracy in responses. Such answers are easy in simpler studies such as micro-tasks. However, as the study grows in complexity towards full scale simulations certain set points can become more ambiguous. During a full-scale control room study, there may be multiple successful paths, some equal to others in efficiency and subjective correctness.

In summation, these three measures are at their best when they are collected automatically and their set points are thoroughly designated. As automaticity and definition decline so does the validity and reliability of the measures. Due to their strengths and simplicity, no further evaluation will be made for such performance measures.

4.2 Workload Measures

4.2.1 Physiological

Physiological measures in general have improved as the technology to collect them has grown smaller, lighter weight, and more sensitive. Such advances have improved their sensitivity to their construct of interest, and reduced obtrusiveness. However, not all methods have continued to improve their value to research.

A spectral analysis of heart-rate variability (HRV) can be used to measure a participant's stress. Stress levels are found in patterns by heart-rate changes or "additional heart-rate" [11]. The stress level is used to interpret workload during different tasks. After a basic literature review it was difficult to determine the true value in this measure.

The strongest measure characteristic may, surprisingly, be unobtrusiveness. Heart rate variability can be captured using a chest band with a couple selectively placed diodes. The minimalist system has little effect on participant's physical and psychological behavior. However, this strength speaks more to the weakness of all the other characteristics than to this device as having potential.

Heart-rate variability is good at measuring stress of the individual but not necessarily workload. It was found that HRV is more sensitive to time-pressure in a task than it is to difficulty of a task [11]. Furthermore, many outside factors must be controlled for such as respiration, muscle activity, body position, physical fitness and age before a baseline measure can be taken [11]. These factors in an individual can also change from day-to-day based on how much sleep or current life stressors are occurring at the time of testing.

Sensitivity of HRV in measuring workload is also low as it failed to detect workload differences in a controlled laboratory setting. It seems only sensitive in testing work/rest differences [12]. It is therefore unreliable, not sensitive to the type of workload which suggests the measures construct validity should be reassessed. HRV does not appear a viable source of information about operator workload and will not be further evaluated in this report.

Functional Near Infra-red (FNIR) measurement, in contrast to HRV, has demonstrated potential across all performance characteristics. It uses light waves emitted from diodes to measure blood flow and

oxygenation levels. The diodes are sensitive only to the area local to each diode allowing different and specific brain structures to be measured at once [13]. It is in this technology that the measurement is appealing. During a workload study involving loading tasks known to require different cognitive resources based on Wickens' Multiple Resource Theory [14] the FNIR was able to differentiate the loaded brain structure during specific tasks. That is, not only was workload measured, but the brain structure performing the work was also detected providing greater insight into the cause of workload. Such sensitivity lends itself to high construct validity. However, care is required when applying the FNIR device to a participant to ensure the desired brain structure is indeed being measured [15]. Finally, workload can be measured throughout the task providing increased resolution. Such sensitivity provides more detailed evidence of the workload source than subjective measures collected post-hoc do.

The FNIR's weakest characteristic is obtrusiveness. The device must be worn at all times and is located on the forehead of the individual. However, this device has been called "negligibly intrusive" by users [15] as it is transportable and does not impair participant actions or senses. It appears, as FNIR becomes more and more accessible its use as a measure for workload should grow.

Eye tracking measures such as pupil dilation and blink rate are also used to assess workload. Pupil dilation is the measurement of pupil diameter and how it changes in relation to different tasks or stimuli [16]. This can be a very reliable type of workload measurement when all of the proper conditions such as ambient light are controlled for. However, it takes effort and time to control for all of these conditions which hurts the overall reliability of this measure [16]. Blink rate is used to assess workload by recording blink frequency within a defined time frame. As the blink rate increases it is inferred that workload was increasing [17].

Eye-tracking technology has improved recently and wearable systems have decreased in size and weight while increasing accuracy and sensitivity. They have become an increasingly reliable source of information. A particular strength is due to eye tracking glasses recording pupil dilation and blink rate in real time during a scenario while the common subjective methods are collected post-hoc, only summarizing the gross workload of the entire scenario [17]. However, eye-tracking measures can still be obtrusive. Participants have reported that there is a period of adjustment in getting used to wearing the glasses throughout a scenario both psychologically and physically [17]. Overall, it has potential as a measure for determining workload.

4.2.2 Subjective Report

One of the most popular measures used is the NASA-TLX. This measurement of workload was developed by NASA and is a multi-dimensional scale that is designed to obtain workload estimates from one or more operators while they are performing a task or immediately afterwards [18].

The NASA-TLX's major strength may well lie in its reliability. As this measure nears universal use, it is becoming effective at delivering useful workload information to researchers [19]. The NASA-TLX is almost completely unobtrusive as well since it is administered at the end of a scenario. However, this is also a weakness in overall sensitivity because workload is only measured at the end of a scenario and not throughout. Furthermore its objectivity is compromised because it relies on participants' own interpretation of the questions in the self-report and perception of their general workload.

4.3 Situation Awareness

There are multiple situation awareness (SA) measurements but only a few that are commonly used in LWRS research. The first is called Situation Awareness Rating Technique (SART). Situation awareness is a rather difficult construct to measure, but SART is the most widely cited measurement of SA [20]. SART provides the participant a tool to rate their own SA. It has a total of 14 components that are generically defined to ensure relevancy to high performance in experiment context. Operators use a series of 2-dimensional scales to rate the degree of their perceived awareness [20].

This measurement is unobtrusive as it is a post-test technique and does not interrupt behavior during a scenario. However, SART shows weakness in objectivity, reliability, and sensitivity because it is a subjective self-report and relies on a subject's opinion which is an indirect measure of SA. Situation awareness must be completely inferred from subjective information. The main appeal behind administering SART is the low cost and ease of use it provides [5].

Another SA measurement is known as Situation Awareness Global Assessment Technique (SAGAT). SAGAT employs freezes in a simulation scenario where operators complete a set of questions that inquire about the current state in the simulation scenario [20]. Freeze probe techniques for measuring SA are generally regarded as objective because they compare participants responses to the actual state in a scenario, rather than relying on a participant's' subjective judgement. Further, situation awareness can vary drastically from one event to another, but SAGAT attempts to account for such variations and establish more reliability. It has strength in sensitivity and as a result is the most popular technique for measuring situation awareness [20]. Researchers at INL have used an adapted version of SAGAT, called SACRI for which they systematically freeze a simulation to measure SA during targeted events. Furthermore, all system information is blocked during a freeze to ensure the participant is not using any displayed information to answer the questions [15]. During this time, a series of queries are provided to the operators to determine his/her knowledge of what was happening at the time of the freeze [20].

The final SA measurement used in LWRS research is eye tracking, specifically fixation points. Fixation points are defined as pauses in which the fovea rests on a particular region of space [21]. This helps researchers determine situation awareness by recording the participant's fixations with eye tracking devices.

The reliability of fixation points depends on the reliability of the equipment. As long as the equipment is functioning properly, fixation points are one of the most useful and reliable measurements in determining where the subjects are looking/fixating [17]. Although reliable eye tracking equipment is objectively measures a participant's visual attention, using that information to infer SA injects subjectivity into the process. SA is usually inferred regarding where a participant *should* be looking compared to where they actually are [22]. A disadvantage to eye tracking is the amount of time required in post-processing to determine a subjects' SA. The process used at INL is especially time consuming as it requires using eye tracking footage to manually map every fixation point onto a reference image.

General measures of SA have room for improvement regarding objectivity and reliability, partially attributed to the difficulty of measuring SA directly. Although freeze-probe techniques and eye tracking directly measure a participant's response, SA must still be inferred from what is gathered and is not a direct measurement. Subjectivity can creep into these measures when trying to characterize what is important in a scenario whether you are measuring visual attention (with eye tracking) or cognition (with freeze-probe techniques).

5 GAP ANALYSIS OF CURRENT METHODS

5.1 Objectivity

Objective measurement is a highly sought after but often not obtained measure characteristic, especially with the number of variables existing in full control room simulation. As defined by NUREG-0711 [6] "A measure should be based on easily observed phenomena" objectivity becomes even more evasive when measuring difficult to observe constructs such as situation awareness and workload. However, advancements in technology have improved the objectivity of some constructs measures.

5.1.1 Workload

The most common measure is NASA-TLX which requires participants to self-report their workload using common descriptions of different workload types and a rating scale. However, the measure still lacks objectivity. When applying workload on a 0-10 scale the value separating zero from one and nine from ten could be different for each participant hence only position on the scale can be used to infer workload and not value.

Some physiological measures are beginning to improve in their capability to measure workload such as eye-trackers and FNIR devices. Both devices have their number of successes. Calculations to measure pupil dilation while accounting for lighting conditions have been improved along with the equipment to track eye movements and measure pupil size. FNIR devices have begun to show promise as it directly measures the blood flow and oxygenation levels of brain structures involved in the processing of the given experimental task. Such direct measure reduces the need for inference increasing workload measure objectivity.

5.1.2 Situation Awareness

The most objective measure of situation awareness involves using eye-tracking to determine a participant's fixation points. The fixation points are then mapped to areas of interest ranked by importance to successful task completion. The measure infers that operators spending more time focusing on areas of importance are therefore more aware of the situation. Although there is still some inference required, the theory applies the eye-mind hypothesis stating that a humans gaze fixation point is indicative of what they are processing [23]. Using fixation tracking technology is the closest strategy found to directly measuring SA, but expert rankings of importance may inject subjectivity into the process.

SA is otherwise measured using participant self-reports and freeze-probe techniques such as SART and SAGAT. Scores on both are found to positively correlate with task performance [24] which is consistent with the theory that good situation awareness is a factor of good performance. Measures like SART cannot claim complete objectivity as self-reports rely on self-awareness and are highly susceptible to individual differences. SAGAT's freeze-probe technique is more objective, but its objectivity may rely on an experts ability to build scenarios and questions sets that capture the important aspects of the environment so that SA can be inferred. Perhaps in the future an even more direct, continuous method may be found to improve unobtrusiveness.

5.2 Construct Validity

Situation Awareness measures were determined to have the most difficult task of measuring construct validity overall. Unlike workload, situation awareness has not found a strong objective measurement tool to begin comparing the results of tools such as SAGAT and SART. Currently, those measures use situation awareness theory to develop questions and compare results to performance, which has been positively correlated. Workload has the fortune of validating new measures with direct measures of processing activity from imaging technology such as fMRIs and EEG to validate. Absence of such resources makes SA a difficult construct to validate.

5.3 Reliability

Reliability is found in a measure that can best control for individual and environmental factors that are outside the experimenter's control. Reliability is also difficult to create in a full scale control room experiment due to the complexity of the task, the interaction between operators, and the diversity of "the correct way" when making system diagnoses. Objective measures are favored during such situations often due to their sensitivity which can therefore attribute measurements to specific times and events. The ability to do so in turn improves the reliability of the measure.

5.3.1 Workload

The NASA-TLX is subject to individual differences and environmental affects outside the control of the experimenter. Under the exact same conditions but on a different day the same participant could reply with different levels depending on factors such as amount of sleep or hunger. Controlling for such factors is unfeasible or impractical. Reliability between participants can be difficult as individual differences in interpretation of "what is workload" can vary from person to person. The construct is relative to each individual making exact replication difficult.

Objective measures did promise some higher level of replicability but have not had the benefit of extensive use like NASA-TLX has had. The fNIR, in part due to its sensitivity, was a reliable measure. Other than experiment design, it was difficult to identify any environmental or individual factors that may strongly influence measurement. Eye-tracking pupil dilation and blink rates can also be more reliable but have individual differences and environmental factors that reduce overall reliability.

5.3.2 Situation Awareness

Subjective reports of SA are susceptible to the same vulnerabilities as NASA-TLX. However, SAGAT is a freeze probe technique targeting a specific population and measures knowledge accuracy of system states as it changes from one situation to another. This improves objectivity however the results rely on the ability of the research team to define the system elements and information that contribute to operator SA.

Measuring SA by eye-tracking fixation to importance ratios is less susceptible to environmental factors and individual differences. Furthermore, due to the measure's sensitivity, data points can be more accurately attributed to specific variables which could increase reliability of the measure.

5.4 Sensitivity

5.4.1 Workload

The NASA-TLX lacks sensitivity due to the way it is most often implemented; at the end of a scenario. The results can only speak to a general estimate of the workload experienced throughout the entire scenario. The result may be an average of the workload experienced throughout the scenario or the highest workload experienced. Either way there is little ability to directly attribute the results to any single event across a complex scenario.

Pupil dilation and blink rates have greater resolution and can be measured during a scenario. The level of workload can be inferred by detecting the extent of dilation and blink rate. Therefore, with continued validation, eye-trackers have a great advantage over the NASA-TLX in detecting small changes in workload and attributing the measurement to specific events. Although the technology has improved from previous versions, it can still lose track of an eye requiring the analysis to be sensitive to such loss in data.

FNIR devices have demonstrated a high-level of sensitivity. Results found FNIR technology was able to detect not only workload levels but was also sensitive to the types of workload tasks taking place by measuring different brain regions at once [15]. FNIR also collects data near continuously during a scenario.

5.4.2 Situation Awareness

SART is employed at the end of a scenario while SAGAT is administered at randomly timed freezes during a scenario. These techniques give researchers a tool to measure SA at strategic points of interest; an improvement toward connecting SA to points in the scenario increasing measure resolution. However,

frequency of data collection is still limited and results are subject to how participants interpret the questions they are asked.

5.5 Unobtrusiveness

Historically there has been a tradeoff between sensitive measures and obtrusive measures. If eye-tracking was to be used, it could only be used for short periods as the weight of the tool would fatigue a participant if left on. However, technology has begun reducing the foot print of such precision tools and increased their mobility allowing greater objectivity, sensitivity and reliability in measures that are minimally obtrusive.

Measuring obtrusiveness is an important consideration during control room studies due to the potential length of a single scenario. Fortunately, measures such as eye-tracking, FNIR and heart-rate have profoundly reduced their obtrusiveness. FNIR for instance, not tested across the length of a scenario was reported as "negligibly intrusive" [15].

Though eye-tracking is a sensitive, objective, reliable, and valid measure of workload, it is the most obtrusive measure. Although current systems weigh approximately the same as a pair of safety glasses, the psychological invasion of privacy may impact operator behavior when using eye tracking technology.

The NASA-TLX is the least obtrusive measure evaluated as it is only performed at the conclusion of a scenario and has zero impact on either psychological or physical behavior of a participant. The SART and SAGAT measures of SA follow suit but are slightly more psychologically obtrusive due to the freeze probe questionnaires. Freeze-probes must be developed with care as not to clue the operator towards collecting or storing information outside their usual patterns.

6 CONCLUSION

The performance measure characterization performed here was intended to find general weaknesses in the battery of performance measures potentially used in an ISV or full-scale simulation study. The purpose was to gain insight as to what measures may need more investigation or development. A secondary goal was to expose some methods, perhaps not often considered, that have improved and offer results in sync with the characteristics suggested by NUREG-0711 [6]. As measures improve they can be integrated in the design process as a full-scale simulation to improve the chances of a successful ISV taking place later. The evaluation performed here was a generalized summary to guide efforts using the 5 performance characteristics as a rubric. The point is to guide future efforts and bring visibility to alternative options.

The three most popular methods discussed here are questionnaires for measuring workload (NASA-TLX) and SA (SART and particularly SAGAT). Their strengths are in their unobtrusiveness and economical implementation. These three measures are also widely used and accepted which can unintentionally translate as gold standard. At the time of their development, other methods (especially physiological methods) were cumbersome, expensive, and less reliable than they are now. Due to the lack of competition the subjective measures often won out and became the ubiquitous go-to that they are today. This paper was completed as a simple reminder that other measures are improving and may require more consideration during experiment design.

Recently, physiological methods have been improving in many ways. Their obtrusiveness has decreased significantly. For instance, eye-trackers are now fully mobile and weigh the same as a pair of safety goggles, although, they can be cost prohibitive. FNIR bands and heart rate monitors are far less obtrusive as well as accessible to smaller budget research. Based on current research it appears these methods, particularly eye-tracking and FNIR meet the performance measure characteristics described in NUREG-0711 [6] more so than the questionnaire and freeze-probe methods do.

7 FUTURE METHOD DEVELOPMENT

Eye-tracking using fixations-of-importance is beginning to show potential as a measure of SA. However, such a measure is thought to only measure level 1 situation awareness [17]. The other measurement options during a full-scale simulation appear to be SAGAT or SART method and could easily be used in conjunction with eye-tracking helping to corroborate findings. Developing methods with increased sensitivity and greater objectivity to measure SA is needed to help ensure proper design evaluation and validation. Further use of the FNIR tool should also be considered. Initial findings show great potential across all five performance characteristics. Most importantly it can provide a continuous measure of workload.

8 ACKNOWLEDGMENTS

This paper was prepared as an account of work sponsored by an agency of the U.S. Government under Contract DE-AC07-051D14517. The views and opinions of the author expressed herein do not necessarily state or reflect those of the U.S. government or any agency thereof.

9 REFERENCES

1. N. (2017). US Nuclear Power Plants. Retrieved February 22, 2017, from http://www.nei.org/Knowledge Center/Nuclear-Statistics/US-Nuclear-Power-Plants
2. Naser, J., Hanes, L., O'Hara, J., Fink, R., Hill, D., & Morris, G. (2004). Guidelines for Control Room Modernization as Part of Instrument and Control Modernization Programs.
3. Boring, R. L., Ulrich, T. A., Joe, J. C., & Lew, R. T. (2015). Guideline for Operational Nuclear Usability and Knowledge Elicitation (GONUKE). Pocedia Manufacturing, 3, 1327-1334. Retrieved February 22, 2017.
4. LeBlanc, K., Boring, R., Joe, J., Hallbet, B., & Thomas, K. (2014). A Research Framework for Demonstrating Benefits of Advanced Control Room Technologies, INL/EXT-14-33901 Revision1
5. Endsley, M.R. (1995b). Measurement of situation awareness in dynamic systems. Human Factors, 37, 65-84. doi: 10.1518/001872095779049499
6. NRC (2012). *Human Factors Engineering Program Review Model* (NUREG-0711, Revision 3). Washington, D.C.: U.S. Nuclear Regulatory Commission
7. Nielsen, J. (1993). Usability Heuristics. Usability Engineering, 115-163. doi:10.1016/b978-0-08-052029-2.50008-5
8. NRC (2002). *Human-System Interface Design Review Guidelines* (NUREG-0700). Washington, D.C.: U.S. Nuclear Regulatory Commission.
9. Hildebrandt, M., & Eitrheim, M. H. (2015). A micro-task method for assessing performance effects of innovative interface elements. Human Factors and Ergonomics Society, 1759-1763. Retrieved February 23, 2017.
10. Spielman, Z., Hill, R., Leblanc, K., Rice, B., Bower, G., Joe, J., & Powers, D. (2016). Full Scale Evaluation of How Task-Based Overview Displays Impact Operator Workload and Situation Awareness When in Emergency Procedure Space. Advances in Intelligent Systems and Computing Advances in Human Factors in Energy: Oil, Gas, Nuclear and Electric Power Industries, 15-27. doi:10.1007/978-3-319-41950-3_2
11. Jorna, P. (1992). Spectral analysis of heart rate and psychological state: A review of its validity as a workload index. Biological Psychology, 34(2-3), 237-257. doi:10.1016/0301-0511(92)90017-o)

12. Nickel, P., & Nachreiner, F. (2003). Sensitivity and Diagnosticity of the 0.1-Hz Component of Heart Rate Variability as an Indicator of Mental Workload. Human Factors: The Journal of the Human Factors and Ergonomics Society, 45(4), 575-590. doi:10.1518/hfes.45.4.575.27094

13. Son, I., Guhe, M., Gray, W. D., Yazici, B., & Schoelles, M. J. (2005). Human performance assessment using fNIR. Biomonitoring for Physiological and Cognitive Performance during Military Operations. doi:10.1117/12.604138

14. Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, 3(2), 159-177. doi:10.1080/14639220210123806

15. Ayaz, H., Shewokis, P. A., Bunce, S., Izzetoglu, K., Willems, B., & Onaral, B. (2012). Optical brain monitoring for operator training and mental workload assessment. NeuroImage, 59(1), 36-47. doi:10.1016/j.neuroimage.2011.06.023

16. Xu, J., Wang, Y., Chen, F., & Choi, E. (2011). Pupillary response based cognitive workload measurement under luminance changes. In Human-Computer Interaction–INTERACT 2011, (pp. 178-185). Springer Berlin Heidelberg. doi:10.1007/978-3-642-23771-3_14

17. Kovesdi, C. R., Rice, B. C., Bower, G. R., Spielman, Z. A., Hill, R. A., & LeBlanc, K. L. (2015). Measuring Human Performance in Simulated Nuclear Power Plant Control Rooms Using Eye Tracking. USDOE Office of Nuclear Energy. doi:10.2172/1261061

18. Hart, S. G. NASA Task Load Index (NASA-TLX); 20 years later

19. Cao, A., Chintamani, K. K., Pandya, A. K., & Ellis, R. D. (2009). NASA TLX: Software for assessing subjective mental workload. Behavior Research Methods, 41(1): 113-117.

20. Endsley, M. R., & Selcon, S. J. (1998). A Comparative Analysis of SAGAT and SART for Evaluations of Situation Awareness.

21. Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. In Proceedings of the 2000 symposium on Eye tracking research & applications, (pp. 71-78). ACM. doi:10.1145/355017.355028

22. Moore, K. & Gugerty, L. (2010). Development of a novel measure of Situation Awareness: The case for eye movement analysis. *Proceedings of the Human Factors and Ergonomics Society*. 54th Annual Meeting. doi: 10.1518/107118110x12829370089788

23. Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4), 329-354. doi:10.1037//0033-295x.87.4.329

24. Durso, F. T., Hackworth, C. A., Truitt, T. R., Crutchfield, J., Nikolic, D., & Manning, C. A. (1999). Situation Awareness as a Predictor of Performance in En Route Air Traffic Controllers.