CONF 890470

FEASIBILITY OF QUANTITATIVE PERFORMANCE MEASURES FOR EVALUATING NUCLEAR POWER PLANT OPERATORS *

Richard J. Carter

Engineering Physics and Mathematics Division
Oak Ridge National Laboratory
Oak Ridge, Tennessee

CONF-890470--2

DE89 007846

Edward M. Connelly

Performance Measurement Associates
Reston, Virginia

Paul A. Krois

Computer Technology Associates
Englewood, Colorado

Presentation To Be Made At:

The American Nuclear Society
Eighth Symposium on the Training of
Nuclear Facility Personnel

Gatlinburg, Tennessee

April 24-27, 1989

* Research sponsored by the U.S. Nuclear Regulatory Commission under U.S. Department of Energy (DOE) Interagency Agreement 40-550-75 with Martin Marietta Energy Systems, Inc. under Contract No. DE-AC05-840R21400 with DOE.

FEASIBILITY OF QUANTITATIVE PERFORMANCE MEASURES FOR EVALUATING NUCLEAR POWER PLANT OPERATORS *

Richard J. Carter
Edward M. Connelly
Paul A. Krois

ABSTRACT

A more valid measure of team performance in nuclear power plants is needed. A study is described which was oriented towards evaluating the feasibility of synthesizing performance measures by deriving measures for crews responding to an off-normal event in a full-scope simulator. The thesis was that performance assessment is based on the subjective judgment of training instructors. The procedure used to synthesize the performance measure consisted of: identification of the factors believed to be important to performance assessment, development of example crew performances and ratings on each be instructors, and derivation of the measure by capturing the instructors' assessment rules. A performance measure was derived which explains nearly all of the variance of the instructors' team performance assessments. There is reason to believe that this method of synthesizing measures can be applied to other events.

INTRODUCTION

The assessment of crew performance in nuclear power plant (NPP) operations usually is accomplished by examining team performance on individual tasks, as well as on the overall exercise. For task performance assessments, relatively global team criteria such as task performance time or number of errors generally are compared to a

---

* The research was sponsored by the U.S. Nuclear Regulatory Commission under U.S. Department of Energy (DOE) interagency agreement 40-550-75 with Martin Marietta Energy Systems, Inc. under contract No. DE-AC05-84R21400 with DOE.

checklist of previously established reference tasks. Such comparisons lead to the specification of the adequacy or inadequacy of crew performance at the task level. With respect to the assessment of team performance on the overall exercise level, specific aggregates of the task criteria typically are utilized. These, in turn, lead to a specification of adequacy or inadequacy of crew performance on the exercise level. Several concerns/problems exist with respect to such performance evaluation attempts. First, there is the concern that the task performance criteria and exercise performance criteria are not valid measures of team performance. Second, there seems to be no evidence in the literature that supports the aggregation of task performance criteria to obtain overall exercise performance criteria. Third, requirements to specify either adequate or inadequate performance dichotomizes the crew performance assessment procedure. Such a treatment tends to ignore important crew performance information concerning the degree to which adequacy or inadequacy are achieved. Because of these and other concerns related to the adequacy of current team performance evaluation methods, there appears to exist a need to define a more valid measure of crew performance.

Oak Ridge National Laboratory (ORNL) conducted a study to evaluate the feasibility of synthesizing performance measures for NPP teams by deriving measures for crews responding to a test problem in a NPP simulator. This paper describes the study and the results therefrom.

METHOD OF SYNTHESIZING PERFORMANCE MEASURES

ORNL defines a team performance measure as a mathematical function of system variables that permits the quantitative evaluation of overall system (team and plant) performance as the crew responds to an off-normal event. The term performance measure is not used to mean data collection (i.e., the recording of values for plant or crew variables); it is not used to denote the performance scores obtained when assessing team performance; and it is not used

to refer to any decisions made after obtaining crew performance scores. Rather, a team performance measure is an analytical statement that incorporates the tradeoffs among the system variables that must be considered in order to assess total crew performance.

What is the source of performance assessment information on team performance; that is, what is the source one can rely on in evaluating crew performance? The study's thesis was that performance assessment is ultimately based on the subjective judgment of the person (or persons) who is accepted as the authority/expert for team performance assessment. This person is assumed to be a training instructor evaluating crew performance when the team is operating the plant (or NPP simulator) in an important exercise such as a certification exam.

The subjective form of an instructor's performance assessment is effectively applied when the instructor evaluates observed performance. To make the performance assessment, the instructor typically makes explicit judgments regarding the relative importance of various crew and system factors to team performance. While making these judgments, the instructor may use quantified variables describing system states, such as: water level, flow rates, and temperatures, and also factors that are not easily quantified, such as: the appropriateness of crew communication, team efficiency in monitoring system parameters, and crew ability to diagnose plant conditions. Thus, training instructors may use their judgment at two levels during the performance assessment process. At one level, the instructors may estimate the value of the factors that are not readily quantified. Then, at the second level, the instructors may use those estimated values along with the values of quantified variables to assess overall crew performance.

While the subjective form of the instructors' performance assessment judgment may serve the instructors well for the evaluation of team activity that the instructors can personally observe, its subjective form makes it difficult for anyone else to similarly evaluate crew performance. Since the problem is the subjective form of the assessment, not the performance assessment itself, it may be beneficial to consider synthesizing a quantitative equivalent to the instructors' performance judgment.

Since instructors are the source of team performance information, the question of how to extract that information must be addressed. The issue is: Should the instructors be asked to write the rules they use to evaluate crew performance or should the instructor's performance preferences be captured as the instructor evaluates observed crew performances? Research on the extraction of information from experts indicates that while experts can successfully demonstrate their judgement ability when working with observed performances, they may not be fully aware of all the factors and tradeoffs they use to produce their judgments. Consequently, the rules that instructors say they use to assess performances should not be relied on as definitive. However, one can ask instructors to provide the rules they use to assess team performance and employ them as a baseline for the performance measure synthesis.

According to the study's thesis, reliable information on crew performance assessment may be obtained by asking instructors to observe and assess team performance. The instructor can be relied on to compare observed performances and indicate a preference for one over another. If one can present descriptions of performances to instructors and have the instructors score them, or at least order them according to performance preference, then the ordering must imply a rule. If that rule can be captured in a quantified form, the mathematical equivalent of the instructor's subjective judgment would be available.

When considering performance assessment of any system, there may be a number of existing measures that one can suggest as candidate performance measures. The question then is: Do these existing performance measures provide the correct assessment of performance? There is a reliable way of comparing any existing measure to the correct measure even though one does not know the correct measure. The key idea is to recognize that the basis for assessment of performance quality is the subjective judgment of an individual (or group of individuals) who is accepted as an authority in assessing performance. The individual or group examines descriptions of possible performances (called performance demonstrations) and concludes that one performance demonstration is preferred over another which, in turn, is preferred over a third, and so forth. This ordering, according to performance preference, defines the performance discrimination task of the correct performance measure. Thus, if an existing measure discriminates performance demonstrations the same way as the experts, there is no reason to reject the measure. On the other hand, if the measure does not discriminate performance as did the experts, it must be rejected and a correct measure synthesized.

PROCEDURE FOR SYNTHESIZING THE PERFORMANCE MEASURE

The procedure used for synthesizing the performance measure in the feasibility study was comprised of the following steps:

1. Identify the factors believed to be important to performance assessment.
2. Develop examples of crew performances and have instructors rate performance on each.
3. Derive the performance measure by capturing the instructors' assessment rules.

Identification of Factors Believed to Be Important to Performance Assessment

Method

Ten Tennessee Valley Authority (TVA) training instructors provided opinions as to what factors should be considered in assessing crew performance. This information was obtained through the use of a questionnaire. The questionnaire consisted of four items, three open-ended and one closed-ended. Question #1 asked for a rating, on a seven-point scale, of the crew as a whole. The instructor also rated the individual crew members, i.e, senior reactor operator, reactor operator, balance-of-plant operator, and shift technical advisor, if applicable. Question #2 was directed at obtaining the factors the instructor believed to be important when assessing crew performance. Question #3 asked for the specific crew actions or behaviors that were assessed as especially good and significantly influenced the instructor performance assessment. Question #4 requested the instructor to provide for the crew those actions or behaviors that could have been improved.

The questionnaires were completed while the training instructors were evaluating twenty-one Watts Bar and Sequoyah NPP crews operating in the TVA simulator located in Chattanooga, Tennessee. The crews were participating in certification and requalification training classes. The situations which were being simulated consisted of five off-normal events (turbine loading, turbine trip, loss of all feedwater, steam generator tube rupture (SGTR), and main steam line break/SGTR/loss of a refuel water storage tank.

Each instructor evaluated at least one exercise (and completed one questionnaire). Training instructor assignment was not a controlled part of the study, so there was variation in the number of exercises evaluated by each instructor. One instructor evaluated twenty training

exercises, while others only three or four. The median number of exercises evaluated (and questionnaire forms completed) by an instructor was eight.

Results

A major difficulty was encountered during the data analysis. It dealt with the instructors use of a variety of different, but apparently synonymous, words to describe the same thing. To limit the effects of an analysis judgment when interpreting (possibly erroneously) the intended meaning of a term, the first data analysis of the completed questionnaires was a sorting of the responses by: crew plant (Watts Bar and Sequoyah), off-normal event (A - E), question, and crew rating (crew rating was classified into three levels - 3.5 points or less, 3.5 to 5.5 points, and 5.5 or more points). An example is presented in Table 1.

The second analysis consisted of investigating the responses to question #2 (factors believed to be important when assessing crew performance). The most frequently occurring terms (or more accurately, the most frequently occurring factors as the analyst perceived them) were: quality of actions, observation/awareness, communication, use of procedures, identification of problem, timely execution, and teamwork. Table 2 gives the results of an analysis of these seven terms. While the answers to question #2 were used to identify the major factors, the data shown in the table were taken from the responses to questions #2, #3, and #4. The percentages (%) shown are the percentages of all answers to the questions that used the factor typed in bold letters.

Discussion

The data generated in this part of the study was sufficient for the purpose of the analysis; however, the analysis was hard to perform and somewhat subjective due to the fact that the: instructors do not use a common set of well-defined terms to describe performance and an individual

Table 1. An Example Sorting of the Questionnaire Responses

Plant: Sequoyah, **Event:** Turbine Loading, **Rating:** 3.5 to 5.5

Question #3 Task Performed Well	Question #4 Task Needing Improvement	Question #1 Crew/Average Rating
SRO - exceptional knowledge of electric, RO - continuously viewed board	Even though SRO knew all about it, procedures (GOI) should have been used more in putting generator on line	5.0/5.4
Use of instructions very good, communication of crew good, start-up accomplished with a minimum of problems, such as no low level alarms on S/G's	Somethings were done by crew members and were not known by others in group, such as transferred station service	5.0/5.0
Communication of group very good	Instructions could have followed more closely, some actions were out of step	5.0/5.0
Good communication in all areas, smooth operation	None - good job done by all	5.0/5.0
RO demonstrated exceptional ability in working with BOP	Supervisor should have communicated more crisply with the operators, BOP should have maintained better control of S/G levels	4.0/4.25
Normal start-up, use of GOI	Getting S/G level control in automatic	4.0/4.2
Crew followed instructions	S/G level control could have been better but problem was simulator, if BOP had practiced on simulator before his performance would have been better. The problem is the speed of the MFP when in automatic	4.0/3.8

Table 2. Results from the Analysis of the Seven Terms

% QUESTION #2	% QUESTION #3	% QUESTION #4	EVENT
	QUALITY OF ACTIONS		
10.5	3.4	25.0	A
16.7	28.6	47.4	B
13.5	16.0	25.0	C
8.3	26.5	17.4	D
16.1	17.4	37.9	E
	OBSERVATION/AWARENESS		
15.8	20.7	6.7	A
0.0	0.0	0.0	B
0.0	16.0	0.0	C
0.0	0.0	4.2	D
0.0	0.0	3.2	E
	COMMUNICATION		
28.9	37.9	21.4	A
25.0	21.4	21.1	B
16.2	20.0	11.1	C
18.8	14.7	9.5	D
16.1	17.4	7.1	E
	USE OF PROCEDURES		
23.7	17.2	14.3	A
16.7	7.1	10.5	B
8.1	8.0	0.0	C
8.3	23.5	8.7	D
9.7	8.7	3.4	E
	IDENTIFICATION OF PROBLEM		
2.6	0.0	32.1	A
8.3	14.3	0.0	B
21.6	20.0	19.4	C
18.8	11.8	8.7	D
38.7	34.8	31.0	E
	TIMELY EXECUTION		
10.5	6.9	17.9	A
0.0	0.0	0.0	B
10.8	12.0	11.1	C
12.5	5.9	8.7	D
0.0	4.2	18.9	E
	TEAMWORK		
2.6	13.8	0.0	A
8.3	0.0	0.0	B
10.8	0.0	0.0	C
8.3	11.1	0.0	D
9.7	8.7	0.0	E

instructor uses various terms to refer to the same factor. In addition, a specific vocabulary for discriminating the level of quality for important crew responses does not exist. For instance, the term "communication" was frequently cited as an important crew function. Yet there are many types of information to be communicated including: statements by crew members of plant problems as they perceive them, statements of hypothesis about the plant problem, announcements of an intended plan of action, announcement of completion of a set of actions, statements identifying an alarm, and asking for advise and information. Also, there are both spoken and unspoken communications among crew members.

For the above reasons, it appears as though a standardized assessment language should be developed. The language should be defined so that it allows instructors to identify particular crew actions and responses. The language should also permit precise descriptions of many levels of response quality that exist between the superior and need-to-be improved performance levels.

Collection of Scored Performance Demonstrations

The purpose of this part of the study was to determine what measurable data (independent variables) will predict instructors' performance assessment scores (dependent variables). It investigated the use of measurements of crew responses which are only available by observation of the crew. According to the study's theory, prediction of the instructor's crew assessment score from independent variables provides the basis for the determination of the quantitative rule for scoring crew performances, i.e., the performance measure sought.

Method

Two experienced NPP training instructors from ORNL evaluated fifteen crew performance demonstrations. They also completed a ten-item questionnaire which was designed

to elicit observational, rather than judgmental, information describing each crew's functioning. The questionnaire consisted of closed-ended items, each having a five-point rating scale; Table 3 presents the ten questions. After the instructors had finished responding to the observational questions, they again rated the crew's performance. The two assessments were collected to determine if the training instructor's thought process of systematically thinking about specific crew responses affects his overall crew assessment. A difference between the first and second assessment scores would indicate the existence of such an effect.

The performance demonstrations consisted of video recordings of the loss of feedwater off-normal event which were taken in concert with the first part of the study. (As crews operated in the simulator, two video cameras recorded their actions and communications.) The video tapes of the NPP crews operating in the TVA simulator contain information on crew interactions, monitoring of the control panel, and crew hypothesis formulation and testing. Each instructor viewed and evaluated fifteen sets of tapes.

Results

It was determined through review of the data that the first and second performance assessments were not always consistent. Instructor 1 changed his assessment four times with two increasing and two decreasing in value. Instructor 2 changed his assessment six times with only one decreasing. While there does not seem to be a pattern to the changes, responding to the observational questionnaire did cause the instructors to reflect on the crew's performance, and in some cases, provide a different assessment.

Consistency between the instructors' scores or the relative ranking of the crews' performance is the first consideration because those judgments are the basis for the

Table 3. Observational Questions

#1. To What Extent Was the Quantity of Communications Proportional to the Rate of Automatic Actions and/or Changes in Important Parameters?				
1	2	3	4	5
Less Than				Greater Than
#2. To What Extent Was the Content of Communications Related to Pertinent Information?				
1	2	3	4	5
Unrelated				Related
#3. To What Extent Was the Style/Quality of Communications Instructions Clear?				
1	2	3	4	5
Unclear				Clear
#4. To What Extent Did the Crew Demonstrate Knowledge of Immediate Actions in Procedures?				
1	2	3	4	5
Few Actions Accomplished				All Actions Accomplished
#5. To What Extent Did the Crew Make Efficient Use of Time Available as a Reflection of Their Knowledge of the Tolerance of the Plant?				
1	2	3	4	5
Delayed Response				Rapid Response
#6. To What Extent Did the Crew Continuously Monitor Parameters?				
1	2	3	4	5
Infrequent				Frequent
#7. To What Extent Did the Crew Continuously Monitor Alarms?				
1	2	3	4	5
Infrequent				Frequent
#8. To What Extent Did the Crew Members Perform Within Their Job Assignments?				
1	2	3	4	5
Not All The Time				All The Time
#9. To What Extent Did the Crew Correctly Diagnose Plant Conditions?				
1	2	3	4	5
No Diagnosis				Complete Diagnosis
#10. To What Extent Was the Crew Continuously Aware of the Plant State?				
1	2	3	4	5
Seldom				Continuously

measures to be developed. Table 4 shows the instructors' scores for each crew, ordered according to score value. As an aid to understanding consistency of instructor scoring, the rankings were divided into quarters, as shown in the fifth column in the table. Seven of the fifteen crews were consistently scored within quarters.

The rank differences of each crew are exhibited in the second column of Table 5. Crews with the largest rank differences were #8 and #12 with a difference of six ranks, while crew #9 had a difference of four ranks. The remaining crew ranking differences were less than four ranks.

In addition to the differences in crew score rankings, differences between the instructors' observations of crew responses, as indicated by the responses to the observational questions, must be considered. A comparison of the instructor observations is provided in the third column of Table 5. It shows for each crew the number of questions with answers different by more than one point. Crews #8 and #12, the crews with the largest differences in performance score rankings, are found to have only one question (question #9) with answers different by more than the tolerance of one.

Table 6 exhibits, for each observation question, the number of crews for which the values of the answers were different by more than one point. The third column of the table indicates the relative prediction power of the questions, which is explained shortly. Examination of Table 6 reveals that question #9 resulted in four crews having answers differing by more than one point. This suggests that question #9 may be poorly stated, resulting in different interpretations by the individual instructors. Perhaps the subject of the question requires a multi-dimensional description, requiring multiple questions. Similar problems may exist, but to a lesser extent, with questions #3 through #7.

Table 4. Relationship Between Performance Scores

Instructor 1 Average Score	Crew	Instructor 2 Average Score	Crew	Consistent In Quarter
95	6	90	6	6
86	3	85	14	14
86	14	82.5	1	
84	8	82.5	3	3
82	15	82.5	13	13
78.5	1	80	2	
77.5	5	80	12	
77	13	77.5	15	15
76.25	2	75	5	
75.5	11	75	8	
73	7	70	9	
65	4	70	4	4
61	10	65	11	
60	12	60	7	
59.5	9	60	10	10

Table 5. Differences in Performance Orderings

Crew	Difference in Ordering of Scores	Number of Observation Questions with Values Different by More Than One Point
12	6	1
8	6	1
9	4	2
7	3	1
11	3	0
2	3	1
13	3	2
1	3	1
15	3	1
3	2	2
10	2	0
5	2	2
14	1	1
6	0	1
4	0	2

Table 6. Differences in Answers to Observation Questions

Question	Number of Crews with Different Answers *	Relative Score Prediction Value
1	0	-
2	0	-
3	2	2
4	2	-
5	2	3
6	2	4
7	3	-
8	1	-
9	4	1
10	1	-

* Number of answers that were different by more than one point.

Table 7. Prediction of Crew Performance

	Mean	Standard Deviation
Dependent Variable: Second Assessment	75.6	9.61
Independent Variable: Question #9	3.7	1.16
Independent Variable: Question #3	3.2	0.83
Independent Variable: Question #5	3.4	0.82
Independent Variable: Question #6	3.5	0.69

Multivariate Regression Analysis

Independent Variable	Regression Coefficient	t-Value	p
Question #9	2.96	4.17	.0001
Question #3	4.94	5.52	.0001
Question #5	3.75	4.23	.0001
Question #6	2.79	2.95	.01

Intercept = 25.57
Variance Explained = 89.4%

A step-wise, multivariate regression analysis was conducted on the data; results are given in Table 7. The dependent variable was the second crew assessment. The independent variables were the responses to the observation questions. Regression results are listed in the order of the variance each explained in a univariate regression analysis. Thus, question #9 explained the most variance and question #6 the least (but still statistically significant). This ordering of the independent variables according to variance explained is indicated in the "relative score prediction value" column of Table 6.

Discussion

Since the variance explained by the regression, which uses only answers to observational questions for independent variables, is very high, strong evidence is available to support the claim that instructors use observations of crew activity, as opposed to simulated power plant data, to formulate their team performance assessments. This might be translated into the conclusion that it is the process the crew employs rather than the results of the crew actions that is of primary importance to instructors when assessing team performance -- but, at this point, this is merely conjecture.

Since, according to the regression analysis, question #9 is important to crew performance prediction and since, in answering that question, the instructors differed as to the crews' ability to diagnose the plant conditions, ORNL must conclude that it is the difference in the instructors' interpretation of question #9 that produced the difference in the crew performance scores. Consequently, the adequacy of the question should be challenged -- instructors should be asked to state their interpretation of the question and expound on their preferences of plant diagnostic strategies. When the different interpretations of the instructors are understood, the question (or perhaps questions) can be

rewritten and the relevant part of the performance assessment questionnaires completed again by the training instructors.

CONCLUSIONS

A crew performance measure was developed which uses, as independent variables, instructor observations of specific crew behavior. Because the measure explains nearly all of the variance of the training instructors' team performance assessments for the loss of feedwater off-normal event, there is strong evidence that the performance measure is functionally equivalent to the subjective assessment rules used by instructors. There is also reason to believe that the method of synthesizing this measure can be successfully applied to other types of off-normal events.

The measure synthesis method permits comparison of differences among instructors as to the relative importance of the affect of crew behaviors on overall system (crew and equipment) performance. Examination of the mathematical function provides a means for identifying the crew behaviors that instructors apparently use in making performance assessments.

RECOMMENDATIONS

Recommendations for future research include the following:

1. The data collection questionnaire should be refined, based on interviews with instructors, to more precisely define the observation variables found to be critical to crew performance assessment.

2. Behavioral examples for each level of each of the observational questions should be developed, so that crew performance corresponding to each level can be more easily understood by other instructors.

3. The performance measurement synthesis should be applied to similar system problems using additional instructors to develop a larger data base of instructor performance assessment judgments.

4. The performance measure synthesis should be applied to additional types of system problems to determine the measurement factors common to all plant problems and those specific to each problem.

5. A standardized performance assessment language and a specific vocabulary for discriminating the level of performance quality should be developed for use by training instructors.

6. Computer programs which can be installed in each NPP facility to automatically synthesize performance measures for any new system problem should be developed.