CONF-9109160--6

PNL-SA-19604

TEAM INTERACTION SKILLS EVALUATION CRITERIA FOR NUCLEAR POWER PLANT CONTROL ROOM OPERATORS

J. C. Montgomery
J. Toquam[a]
C. Gaddy[b]

September 1991

Presented at the
Human Factors Society Annual Meeting
September 2-6, 1991
San Francisco, California

Prepared for
the U.S. Department of Energy
under Contract DE-AC06-76RLO 1830

Pacific Northwest Laboratory
Richland, Washington 99352

(a) Battelle Human Affairs Research Center, Seattle, Washington
(b) General Physics Corporation, Columbia, Maryland

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

TEAM INTERACTION SKILLS EVALUATION CRITERIA FOR NUCLEAR POWER PLANT CONTROL ROOM OPERATORS

Joseph C. Montgomery
Battelle Pacific Northwest
Laboratories
Richland, Washington

Catherine Gaddy
General Physics Corp.
Columbia, Maryland

Jody Toquam
Battelle Human Affairs
Research Center
Seattle, Washington

Abstract

Previous research has shown the value of good team interaction skills to group performance, yet little progress has been made on in terms of how such skills can be measured. In this study rating scales developed previously (Montgomery, et al., 1990) were extensively revised and cast into a Behaviorally Anchored Rating Scale (BARS) and a Behavioral Frequency format. Rating data were collected using 13 training instructors at the Diablo Canyon Nuclear Plant, who rated three videotapes of simulator scenario performance during a day-long training session and later evaluated control room crews during requalification training. High levels of interrater agreement on both rating scales were found. However, the factor structure of the ratings was generally inconsistent with that hypothesized. Analysis of training ratings using Cronbach's components of accuracy (Cronbach, 1955) indicated that BARS ratings generally exhibited less error than did the Behavioral Frequency ratings. The results are discussed in terms of both field and research implications.

Introduction

Team interaction skills include the way that members are able to "pool their abilities in a collaborative context in order to reach the best decision" (Watson & Michaelsen, 1988). Such collaboration may involve sending clear messages, attending to what was said, questioning what was not understood, providing useful suggestions and ideas when needed, participating in group decisions, resolving conflicts between crew members, managing stressful situations, providing assistance to overworked fellow crew members, and providing an emotional investment in the group's activities.

Unfortunately, reliable, valid measures of team interaction skills are still in need of development (Dyer, 1984). Recent work by Komaki, Desselles, and Bowman (1989) obtained inter-rater agreement exceeding 95% in a study of crew behavior in sailboat races. Team skills assessment has also been the focus of ongoing programs sponsored by the Naval Training Systems Center (Morgan, Glickman, Woodward, Blaiwes, & Salas, 1986; Glickman, Zimmer, Montero, Guerette, Campbell, Morgan, & Salas, 1987; Oser, McCallum, Salas, & Morgan, 1989). In each of these studies, better team skills was associated with higher levels of overall team performance. Research on team skills has also been in progress in commercial and military aviation as well as in the aerospace area (Foushee, 1984; Orlady & Foushee, 1986). However, to date

little research has addressed the issue of how team interaction skills can be measured in the context of nuclear power plant control room operations. The present study served to develop and evaluate such measures, focusing on team performance in simulated emergency conditions.

Method

Rating Scale Development

The rating scales developed in Montgomery et al. (1990) were thoroughly revised prior to additional data collection efforts. These scales, developed on the basis of a literature review and a scale development workshop involving contract license examiners, included a Behaviorally Anchored Rating scale (BARS) and a Behavioral Frequency scale format, both based on the same seven dimensions.

Evaluation of data collected using these rating scales indicated difficulty in the areas of interrater reliability, obtained factor structure, and ability of control room personnel to use three of the dimensions. In the revision process the three problematic dimensions were eliminated, the remaining dimensions simplified, and two new dimensions were introduced. The resulting dimensions closely followed the factor structure produced by exploratory factor analyses. For the BARS, a seven point scale was again used, with behavioral anchors developed for "High" "Medium" and "Low" performance. In addition, an extensive revision of the Behavioral Frequency scale items was undertaken. These items relied on a seven point scale as well, ranging from "Always" to "Never". The following six dimensions emerged from the revision process and were used in this study: Communications, Task Coordination, Team Spirit, Maintaining Task Focus in Transition, and Adaptability.

Subjects

The participants in the study were 13 Instructors and Senior Instructors at the Diablo Canyon Nuclear Plant. All possessed a Senior Reactor Operator license and averaged 12.4 years of nuclear operating experience. All had at least a high school education, but over half possessed AA or BS degrees.

Training

All subjects participated in a day-long training session presented by three research team members. The training included discussions of the meaning and importance of team interaction skills, definitions of the dimensions of team interaction skills used in the study, and familiarization with the BARS and Behavioral Frequency scale formats. In addition, participants viewed three videotape scenarios, each lasting approximately ten minutes, which depicted a control room crew displaying poor, high, and moderate levels of team interaction skills. After each scenario, participants rated the team interaction skills displayed, then participated in a discussion of the ratings. The training format was highly interactive throughout the day and appeared to be very well received by the participants.

Data Collection

Primary data collection consisted of three research team members and several of the training instructors observing control room crew performance in the Diablo Canyon simulator during requalification training. A total of eight crews, each responding to three simulator scenarios, were observed. Each scenario lasted from one to two hours and required the crews to respond to a simulated plant emergency. Following each scenario, research team members and the participating training instructors rated team interaction skills observed.

The ratings of the videotape scenarios obtained during training served as a secondary data source. Since "true scores" were available for these scenarios (the ratings provided by the research team, who had developed the tapes) it was possible to compute Cronbach's components of accuracy (Cronbach, 1955) in order to compare rating errors on the two rating scales and to examine trends in errors across the three scenarios.

Results

Descriptive Statistics

Ratings of control room crew simulator performance revealed little difference by scenario, and were combined for computation of descriptive statistics (see Table 1). Means for the BARS ratings all exceeded 5.3 (on a 7-point scale), with relatively small standard deviation values. Similarly, mean values for the Behavioral Frequency scales were quite high (minimum of 4.57), with small standard deviations. Ratings on both scales thus demonstrated a considerable degree of negative skew.

Table 1
Descriptive Statistics for BARS and Behavioral Frequency Ratings

	BARS Ratings			Behavioral Frequency Ratings		
Dimension	Mean	S.D.	N	Mean	S.D.	N
1. Communications	5.38	.92	176	5.47	.67	176
2. Openness	5.51	.82	176	4.57	1.14	176
3. Task Coordination	5.37	1.02	175	4.76	1.18	177
4. Team Spirit	5.42	.95	175	4.83	.97	177
5. Maintaining Task Focus	5.51	.87	175	4.68	.92	177
6. Adaptability	5.55	.87	174	5.35	.84	176

Interrater Reliability

Assessment of interrater reliability was complicated by the extreme degree of negative skew and range restriction present in the ratings. James (1984) argued that the use of the Pearson correlation or the intraclass correlation is inappropriate in such situations. Both approaches can provide spuriously low or even negative values, even if ratings across raters are nearly identical. The approach devised by James (1984) is essentially based on comparing obtained rating variance with variance that might result from purely random ratings or from random ratings taken from moderate to extremely skewed distributions. That is, ratings might contain no true score variance,

but reflect leniency or a social desirability response bias that would reduce variance and skew the distribution. Such response biases might give at least the appearance of interrater agreement (such as all high ratings) where no agreement actually existed. The James (1984) approach thus assesses interrater reliabililty while compensating for potentially spurious agreement.

Computation of interrater reliability for each crew and scenario, based on the James (1984) approach, resulted in consistently high values, whether comparisons were based on random, moderately skewed, or extremely skewed hypothetical distributions. For the most conservative case, comparison with an extremely skewed distribution, the average interrater reliability for the BARS was 0.89 (ranging from 0.72 to 0.99. For the Behavioral Frequency ratings, the mean interrater reliability was 0.86 (ranging from 0.55 to 0.96). (Note that for each rating scale type, a total of 24 reliability coefficients were computed, given eight crews each performing three scenarios.)

Confirmatory Factor Analysis

Confirmatory factor analyses of both the BARS and Behavioral Frequency ratings were completed using the EQS structural equations program (Bentler, 1989). For the BARS, a single-factor solution proved to be a reasonably good solution (normed fit index=.92, chi-squared=48.35, $p<.001$), although the best fit was a two-factor solution, with cross-loading of one dimension (normed fit index=.98, chi-squared=15.2, NS). That is, Communications, Openness, Task Coordination, and Team Spirit formed one factor and Task Coordination, Maintaining Task Focus, and Adaptability formed a second, with Task Coordination loading on both factors. The hypothesis that all six dimensions were independent received little support.

For the Behavioral Frequency ratings, the six factor model was also a poor fit with the data (normed fit index=.39, chi-squared=1388.05, $p<.001$). The best model was a five-factor oblique solution (normed fit index=.73, chi-squared=620.02, $p<.001$). An exploratory factor analysis was also conducted, resulting in a five factor solution. The individual Behavioral Frequency items, in general, did not load consistently on their hypothesized dimensions.

Internal Consistency Reliability

Coefficient alpha was computed for the BARS ratings as a whole, for the Behavioral Frequency scales as a whole, and for the individual dimensions within the Behavioral Frequency Scales (each of which contained four or more items). Coefficient alpha for the BARS was 0.902, and for the Behavioral Frequency ratings overall was 0.812. On individual dimension basis within the Behavioral Frequency scale, alpha for Communications was .68, for Openness .44, for Task Coordination .54, for Team Spirit .48, for Maintaining Task Focus .25, and for Adaptability .52.

Training Ratings

Analyses were also performed of the ratings provided by study participants after viewing each of the three videotape scenarios (which depicted poor, high, and moderate performance, respectively). Table 2 contains mean ratings, across dimensions and raters, for each rating scale and for each scenario as well as "true score" ratings provided by the research

team. This comparison provides a rough sense of the match between true scores and actual ratings. As can be seen, the average ratings closely matched the true score ratings for the BARS. For the Behavioral Frequency ratings, however, the mean showed little deflection from the center of the seven point scale regardless of whether the scenario depicted high or low performance.

Table 2
Comparison of Mean BARS and Behavioral Frequency Ratings with True Scores

	Mean BARS Rating		Mean Behavioral Frequency Rating	
	Raters	True Scores	Raters	True Scores
Scenario 1	2.06	2.67	3.79	2.69
Scenario 2	6.33	6.5	4.08	4.96
Scenario 3	4.46	4.17	3.97	4.09

Cronbach's components of accuracy (Cronbach, 1955) were computed to further investigate sources of rating error. The four accuracy components consist of elevation (tendency of an individual to rate above or below the average true score), differential elevation (tendency to rate a given scenario, across dimensions, higher or lower than is warranted), stereotype accuracy (in which an individual's ratings of a particular dimension may be consistently too high or low), and differential accuracy (rater error left after controlling for overall, scenario, and dimension effects). The four accuracy components were calculated for each type of rating scale, resulting in eight accuracy scores for each rater. In addition, two of the components (elevation and stereotype accuracy) were computed for each scenario. (The other two accuracy components cannot be computed on a scenario-wise basis.) It should be noted that higher values of an accuracy component indicate the presence of greater error, not more accuracy.

Paired t-tests were then computed, comparing the BARS accuracy components with the corresponding Behavioral Frequency accuracy components. Elevation for the BARS exceeded that of the Behavioral Frequency ratings (t= 5.5, p <.0002), while differential elevation was greater for the Behavioral Frequency ratings (t= -5.01, p <.0004), as was differential accuracy (t= -3.8, p <.003).

The two accuracy components that were computed on a scenario-wise basis were each entered into an analysis of variance design to examine scenario, type of rating, and scenario by type of rating interaction effects. For elevation a significant scenario effect was found, $F(2,55)=16.42$, p <.00001 as was an interaction effect, $F(2,55)=13.55$, p < .00001. A simple effects analysis determined that elevation for the Behavioral Frequencies decreased significantly with each scenario, while no significant differences were obtained across scenario for the BARS. A similar analysis of variance was performed using stereotype accuracy. For this component a significant scenario effect was found, $F(2,55)=49.9$, p <.00001. A Student-Newman-Keuls test indicated a significant reduction in error with each scenario. In addition, a rating scale effect was found, $F(1,55)=11.64$, p <.001, indicating that stereotype accuracy error was significantly greater on the Behavioral Frequency scales than on the BARS.

Discussion

The goal of the present study was to develop reliable, valid measures of team interaction skills. Both the BARS and the Behavioral Frequency scales were found to demonstrate a high degree of interrater reliability. Without such solid interrater agreement, it would make little sense to proceed to use the scales, since raters would be likely to make their ratings in an ideosyncratic fashion.

Less successful was the intention that the six dimensions used in each scale would prove to be orthogonal, or relatively independent measures of an aspect of team interaction skills. However, such was not the case. For the BARS, two factors were found, one relating more to interpersonal issues and a second factor more directly task-related. In fact, intercorrelations among the six dimensions were quite high and a one-factor solution provided a reasonably good fit to the data. For the Behavioral Frequency ratings, a five factor solution was found to fit the data, but the individual items failed to load on dimensions as predicted. Thus, the BARS scale might be used to generate either two team interaction skills scores, or (given its high internal consistency) might be considered to be composed of essentially parallel items, and a single score generated from ratings. However, use of the Behavioral Frequency scale appears problematic at this time.

The training data provided further support for the usefulness and validity of the BARS ratings. For example, average BARS ratings closely matched true score ratings. That is, raters successfully used high, medium, and low points on the seven-point BARS rating scales to reflect high, medium, and low scenario performance. For the Behavioral Frequency ratings, however, relatively little dispersion around the scale midpoint was found, regardless of the level of scenario performance.

Analysis of Cronbach's components of accuracy also supported use of the BARS. The BARS ratings contained less rating error due to differential elevation and to differential accuracy as compared to the Behavioral Frequency ratings. Although elevation error appeared to be lower in the Behavioral Frequency ratings, this finding may have been something of an artifact, since computation of elevation involves the difference between average ratings compared to average true score ratings. Average true score ratings for the low, high, and moderate performance scenarios fell very close to the scale midpoint, as did the average Behavioral Frequency ratings for eacha of the three scenarios, resulting in very low levels of elevation.

The scenario-wise analysis of Cronbach's components showed that the stereotype accuracy error in both BARS and Behavioral Frequency ratings decreased across training. This finding supports the use of an interactive training model, with feedback provided on ratings, as a way to improve accuracy of ratings. In addition, it was found that the stereotype accuracy error for the BARS was significantly lower than that for Behavioral Frequency ratings for each of the three scenarios.

Additional research is needed to discover why it should be the case that the BARS appeared to be more successful at measuring team interaction skills than were the Behavioral Frequency ratings. Theoretically, BARS ratings require complex judgments, while frequency ratings rely on a simpler recall

process. However, research by Murphy, Martin, and Garcia (1982) has suggested that frequency ratings do not simply measure observations or direct recall, but are "disguised measures of traitlike judgments" (p. 566). Thus the rating process may be quite similar for both scales. In addition, Gaugler and Thornton (1989) found that assessment center judgments based on only a few dimensions were more accurate and displayed less method bias than ratings based on a larger number of dimensions. Given the relatively large numbers of items rated on the Behavioral Frequency scales (22) as opposed to the BARS (six dimensions), it may simply be the case that the Behavioral Frequency scales imposed too high a cognitive demand on the raters, and that greater error in the Behavioral Frequency ratings is only to be expected.

Acknowledgements

This research was supported by the Human Factors Branch, Office of Nuclear Regulatory Research, U.S. Nuclear Regulatory Commission under Contract DE-AC06-76 RLO 1830. We gratefully acknowledge the support of the NRC Project Manager, Joel Kramer, as well as the assistance of the Diablo Canyon Nuclear Plant training staff and control room crews for their participation in the study. We also thank the training staff of the Limerick Generating Station for their assistance in developing the training videotape.

REFERENCES

Cronbach, L.J. (1955). Processes affecting scores on "understanding of others" and "assumed similarity". Psychological Bulletin, 90, 218-244.

Bentler, P.M. (1989). EQS Structural Equations Program Manual. Los Angeles,CA: BMDP Statistical Software Inc.

Dyer, J.L. (1984). Team research and team training: A state-of-the-art review. Human Factors Review, 285-323.

Foushee, H.C., (1984). Dyads and triads at 35,000 feet: Factors affecting group process and aircrew performance. American Psychologist, 39, 885-893.

Gaugler, B. B. & Thornton, G.C. III. (1989). Number of assessment center dimensions as a determinant of assessor accuracy. Journal of Applied Psychology, 74(4), 611-618.

James, L.R., Demaree, R.G., and Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. Journal of Applied Psychology, 69(1), 85-98.

Komaki, J.L., Desselles, M.L., & Bowman, E.D. (1989). Definitely not a breeze: Extending an operant model of effective supervision to teams. Journal of Applied Psychology, 74, 522-529.

Montgomery, J.C., Gaddy, C.D., Holmes, C.W., Seaver, D.A., Haut, J.T., Spurgin, A.J., Beare, A.N. (1990). Team Skills Evaluation Criteria for Nuclear Power Plant Control Room Crews (PNL-7250). Richland, WA: Battelle Pacific Northwest Laboratories.

Morgan, B.B., Glickman, A.S., Woodard, E.A., Blaiwes, A.S., & Salas, E. (1986) Measurement of Team Behaviors in a Navy Environment. NTSC TR-86-014. Orlando, Fl: Naval Training Systems Center.

Murphy, K.R., Martin, C., & Garcia, M. (1982). Do behavioral observation scales measures observation? Journal of Applied Psychology, 67(5), 562-567.

Orlady, H.W., & Foushee, H.C., (1986). Cockpit resource management training. NASA Conference Publication 2455. Washington, DC: NASA.

Oser, R., McCallum, G.A., Salas, E., & Morgan, B.B. (1989). Toward a definition of teamwork: an analysis of critical team behaviors. NTSC TR-89-004. Orlando, Fl: Naval Training Systems Center.

Shiraki, C. (1989, January). Operator Licensing Examiner Standards. Report No. NUREG-1021 (Revision 5). Washington, DC: Nuclear Regulatory Commission.

Watson, W.E. & Michaelsen, L.K. (1988). Group interaction behaviors that affect group performance on an intellective task. Group and Organization Studies, 13(9), 495-516.