TECHNIQUES OF QUANTITATIVE PERFORMANCE MEASUREMENT FOR ASW TEAM TRAINERS

BERNARD W. YAEGER and JAMES D. BELL
Honeywell Marine Systems Division
California Center

INTRODUCTION

The Need for Quantitative Performance Measurement

Rapidly escalating costs in operating weapon systems, coupled with energy shortages, has drastically affected at-sea combat team readiness training. Simulator/trainers are increasingly depended upon as the only economically effective alternative to provide this training. The increasing dependence on simulators, along with the availability of advanced computer technology, has led to higher demands for proof of training effectiveness. Team trainers, in particular, due to problem complexity, need validation of effectiveness. This need for proven effectiveness has placed an unparalled demand on the instructor's ability to objectively and comprehensively evaluate individual subteam and team performance.

Evaluation of complex interrelated and interdependent performances in team exercises is extremely difficult at best. This difficulty has resulted in elevated interest in the application of CAI/CMI. But the use of CAI techniques requires the capability to effectively evaluate the current performance level of trainees as well. (Goldstein, 1974). This evaluation is necessary so that appropriate courseware logic can be brought to bear as the driving function in the control of exercise content, complexity, and difficulty. Therefore, the development of objective performance measurement techniques is a prerequisite to the use of CAI as well as for proof of training effectiveness.

Honeywell is currently involved in a study of techniques and concepts that will provide detailed team, subteam, and individual performance measures for training evaluation. The purpose of this paper is to describe the requirements for ASW performance measurement and a preliminary concept of a model which meets these requirements.

Background - State of the Art in Performance Measurement

It was noted at the 9th NTEC Industry Conference, 1976 that:

"After more than 30 years, the rating scale remains the basic evaluation method for instructors" (Shipley, Hagin, and Gerlach, 1976).

It was further noted, with respect to ASW trainers specifically, that ASW exercise evaluation relies primarily on subjective instructor scoring of loosely defined areas of performance. Figure 1 illustrates this scoring method and points up the conclusion that ... "A valid scoring system requires the use of objective methods ..." (Copperman and Asa-Dorian, 1976). This example draws attention to the fact that current evaluation: (a) is heavily dependent on subjective opinion, (b) places heavy emphasis on communications evaluation, and (c) is not efficient in identifying specific areas of needed improvement, with examples of poor performance. Clearly, the need for valid and reliable quantitative performance measurement paradigms is mandated by the need for proof of effectiveness and applications of CAI.

A previous unpublished study found that as many as 60 to 70 measures of performance were available and could be used for ASW training. However, the usefulness of measures simply because data are available to obtain them, is questionable. Vruels and Goldstein (1976) have correctly admonished users against the unsystematic application of performance measures. They point out that measures must be selected which: (a) tend to eliminate redundant information, (b) are sensitive to skill changes which occur during training, and, (c) have performance predictive qualities. They presented a method for selecting measures based on multivariate statistical models, which evaluate the total set of candidate measures.

*In addition to the problem of unsystem-*atic application of measures, another misstep that must be avoided is the application of approaches developed for other purposes. Most previously developed or proposed models simply apply sets of measures. Sophisticated approaches to a selection of measures, and statistical reduction of data have been developed, yet the evaluation approach itself remains basically unstructured. (References a, c, d, and f). Two major obstacles are manifest in these approaches. First, the measures that are available tend to be directed

COMMAND/EVALUATOR

AVAIL.	EARNED	
		COMMAND/EVALUATOR - GENERAL
10		Was SAU Commander aware of tactical employment of all units under his command?
10		Was a low noise level maintained in CIC?
10		Was classification a continuous process by SAU Commander?
20		Were four weapon attacks successful? (An urgent attack does not need to be a hit in order to be considered successful.)
		APPROACH PHASE
8		Was Datum properly disseminated to all units?
6		Was SAU properly formed up, and was spacing correct, considering predicted sonar range?
8		Was SAU search front properly re-ordered and was approach to Datum proper for tactical situation?
6		Were Cone of Courses, Intercept Course, and time to enter TDA compared and concurred with Assist Ship?
6		Were appropriate countermeasures executed during Approach Phase?
8		Were Plans Red and Black and Weapons Policy passed to Assist Ship?
8		Was controlling station properly kept informed?
6		Were aircraft plots evaluated to determine target course and speed?
8		Was time to enter TDA updated with latest contact information and new Intercept Course executed accordingly?
5		SWAP SITREP requested and obtained prior to execution of SWAP.
5		SWAP SITREP disseminated to command, with State, Weapons, Contact status, and Datum information included.
7		SWAP executed in a timely manner.
5		SAU advised when SAC is assumed.
8		Zig-Zag Plan executed prior to entering TDA.
6		Appropriate Material Countermeasures prepped or executed.

Figure 1. ASW Escort Qualification Program Evaluation Form

towards measure of individual skills. The interpretation of such measures with respect to tasks which are interrelated between groups of individuals is elusive. Secondly, previous models tend to assume the training exercise in question is a complex integrated and continuous task. Therefore, measures are selected (usually for post exercise critique) which attempt to summarize aspects of performance throughout the entirety of the exercise. The danger of such an approach has been incisively described (Vruels and Goldstein, 1976).

> "Measures which are not useful for one condition, but which are 'carried along' to cover a second condition, might degrade the power of the set to describe the first condition. Thus, one must be cautious in the application of universal measure sets to cover a variety of task situations."

The unsatisfactory applications of previous approaches illustrate that performance measurement methodology for team training needs more development. For this reason, we must first start by addressing the requirements for ASW team training.

APPROACH

Requirements for ASW Performance Measurement

ASW team training exercises such as those using the 14A2 trainer, train several subteams, including CIC, UB Plot, Sonar and Air Control. Training is directed at the development of coordination of operational procedures, and communications both within and between subteams. Tactics is taught at a team level. Different exercises address divergent training objectives, such as ASROC Weapon Delivery, ASW Escort Missions, and ASW Helo Vectored Attack exercises. These exercises, in turn, may be divided into discrete phases such as: *Approach*, *Search*, *Tracking*, *Attack*, *Post Attack*, and Lost Contact. Phases commence and are terminated by key events, such as contact, weapon assignment, etc., which define the conditions within a phase. Task requirements can be significantly different in various phases. Significant variables effecting performance, in ASW training, include environmental conditions, complexity of target submarine maneuvers, and (simulated) malfunctions. The team must be exposed to and trained to handle these varying conditions.

The complexity of the ASW combat readiness precludes a more comprehensive discussion of the training problem within the scope of this paper. However, the requirements for *performance measurement* can be briefly summarized as follows:

1. The set, or composite of quantitative performance measures must be comprehensive, and based on observed or computer recorded data (qualitative data should be avoided).

2. The set of measures must provide valid contributions to team performance evaluation (sensitive to performance change).

3. The set must define performance in a manner such that specific performance needing improvement can be identified (avoid loosely defined parameters).

4. The set must be readily modifiable to reflect the training objectives of different types of exercises.

5. The set must include performance measures from all phases of the exercise.

6. The set must include measures of team, subteam, and individual performance.

7. The set must be adaptable to different levels of complexity in several exercise variables.

8. The set of measures must be subject to weighting which reflects the user's desired emphasis of the various factors of performance.

Preliminary Concept for an ASW Team Performance Measurement Model

Attempting to devise an approach for a Performance Measurement Model that meets the requirements previously discussed above, has led Honeywell to adopt a three-dimensional matrix model as illustrated in Figure 2. The first dimension represents individuals, which in turn, are grouped into subteam blocks. The *second dimension represents categories of* measurement, which includes: (a) procedures, (b) communications, (c) accuracy, (d) tactics and, (e) time. The results of our analysis indicates that most meaningful measures of ASW team training fall within one of these categories. The third dimension represents phases of the training exercises. Any specific exercise may consist of one or more phases depending upon training objectives. Thus, performance measures are referenced to specific subteam/phase/category problem components which are designated as cells.

Each cell limits the area of measurement to a specific operational definition. Therefore, very specific quantitative performance measures can be developed with highly definitive meaning and interpretation. Each

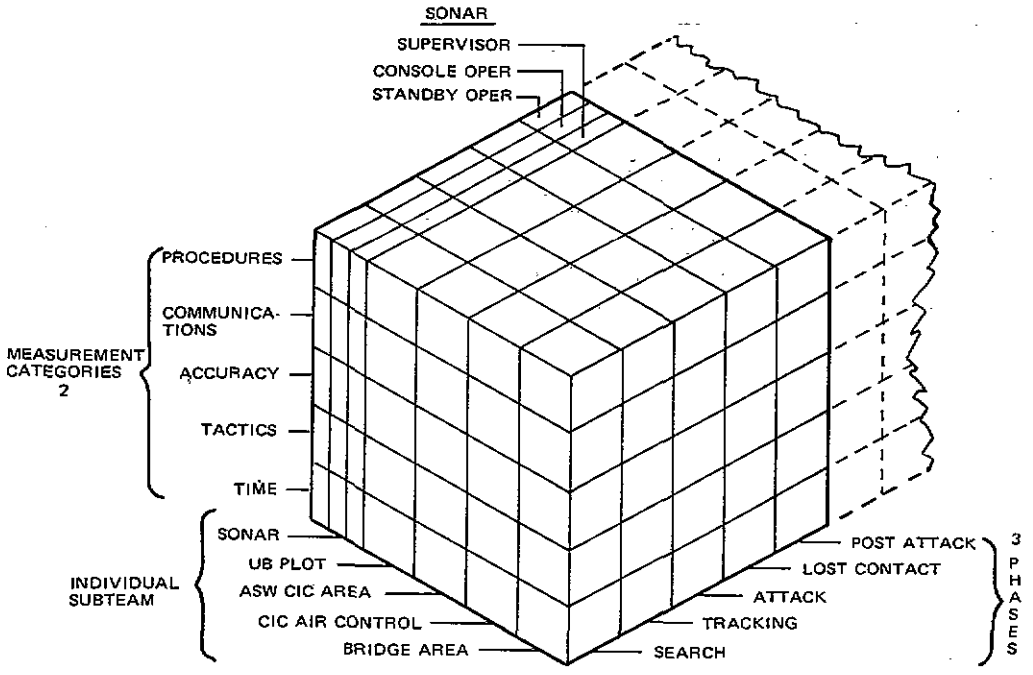

Figure 2. Three-Dimensional Matrix Performance of Measurement Model

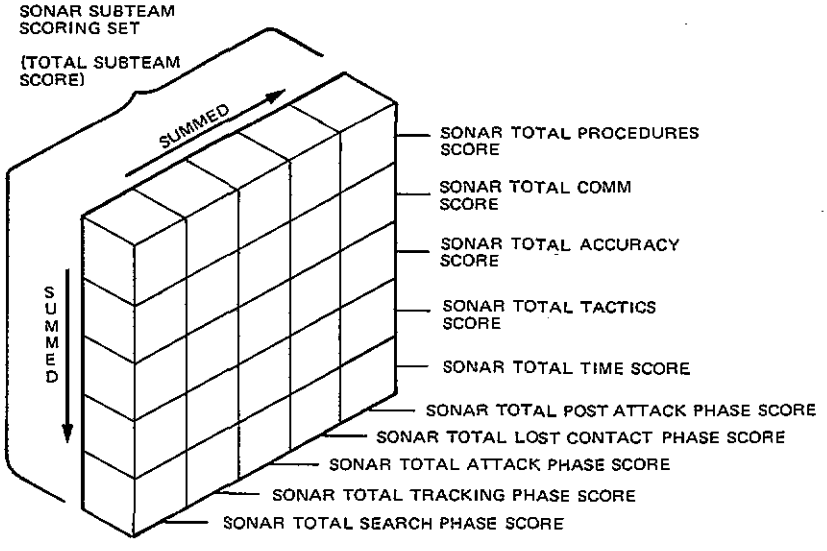

Figure 3. Individual Subteam Scoring Composite

measure addresses a specific performance in a specific phase or task condition. While it is not necessary or even intended that a quantitative measure can be developed for every cell, partitioning into cells enables many measures to be used which would be inappropriate if they were interpreted as measures of total problem, total phase, or total subteam performance.

Combining Performance Measures

Individual performance measures can be used for feedback to the trainees, but they cannot be directly combined to represent performance at a broader level of performance evaluation. Given the quantitative measures for all appropriate cells, composites of these measures must be summated to provide an evaluation by categories, phases, and personnel levels. This summation is possible when individual measures are transformed or converted into standard scale scores ("Z" scores, for example). The total of all individual cell scores represents the total team score. Although the total score summarizes the total team performance level, it is not sufficiently descriptive.

The set of individual cells may be combined into three major composite groups. These composites are:

1. Subteam Composite Scores
2. Exercise Phase Composite Scores
3. Measurement Category Composite Scores

Figure 3 shows an example of the composite of cells for the sonar subteam score. This score represents subteam performance across all phases and categories of measurement. The total composite can be divided into subsets by summing across rows or across columns. Scores summed across rows provide subteam scores by measurement category, i.e., Sonar Procedures Score, for all phases. Scores summed across columns generate subteam scores in all categories, for a specific phase, i.e., Sonar Attack Phase Score.

An example of a Composite score by Exercise Phase is illustrated in Figure 4. The Phase Composite Scores represents the total team score for the designated phase taking into account all subteams and categories of measurement. Subsets of scores summed in rows results in a team score for each category with respect to the phase represented by the composite. Summation of the column subset of scores gives subteam scores for all categories within that phase.

An example of the Measurement Category Composite Score is depicted in Figure 5. The composite is the team score for all phases in the procedure category of measurement. Examination will show that subsets of scores within this composite yield redundant information obtained by the subsets of scores previously described.

The capability of the model to provide a set of scores by column and row, a composite score by subteam, category or phase and a total score for all cells offers a unique opportunity to evaluate team performance and select new problems. For example, if a team is weak in communication, during the lost contact and attack phase, a problem can be selected (or modified) to provide extra practice in these phases with team concentration in improving communication. Similarly, if the team is weak in tactical coordination a problem with this emphasis can be selected. In general, the proposed methodological approach can provide an average comparative score and part scores showing the particular strength and weaknesses of each team.

Measurement Techniques

The measurement of performance for each cell described above depends upon the ability to detect the performance involved. Obviously the computer can measure all performance results which interact with the program, but many other very desirable and necessary performance measures are not available to the computer. These team performances must be observed and measured in some way by an instructor or problem monitor. The following discussion identifies the computer versus monitor requirement for each performance category.

- Performance Monitoring

  One of the primary objectives of ASW team training is teaching trainees to follow and practice the established procedures for their tasks. The following of established procedures or doctrine helps to assure effective team performance. Procedure following, as considered here, consists of console and equipment operations, manual operations involved in plotting, status board updating and log keeping. Procedures in conducting communications are covered in the separate communications category.

  For the purposes of evaluation and training, all incorrect procedures should be detected so each occurrence can be recorded and weighted for criticality. Feedback to the team is necessary for correction.

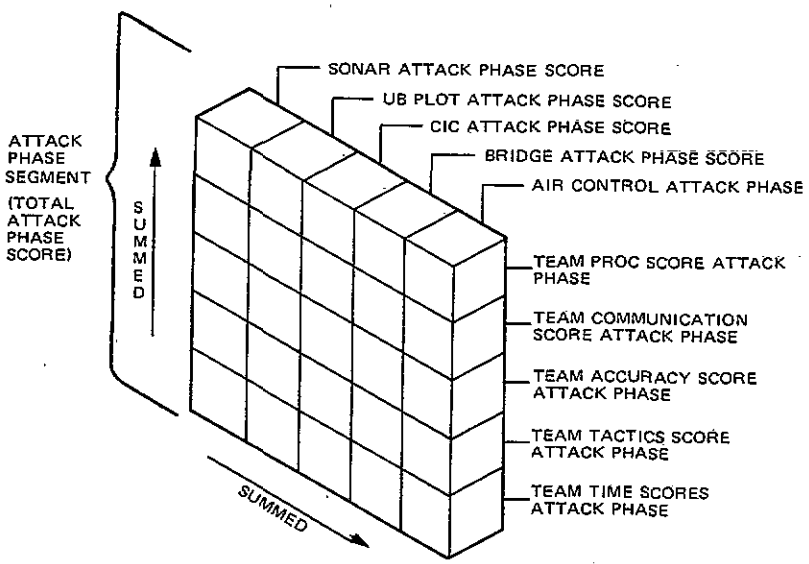

Figure 4. Exercise Phase Scoring Composite

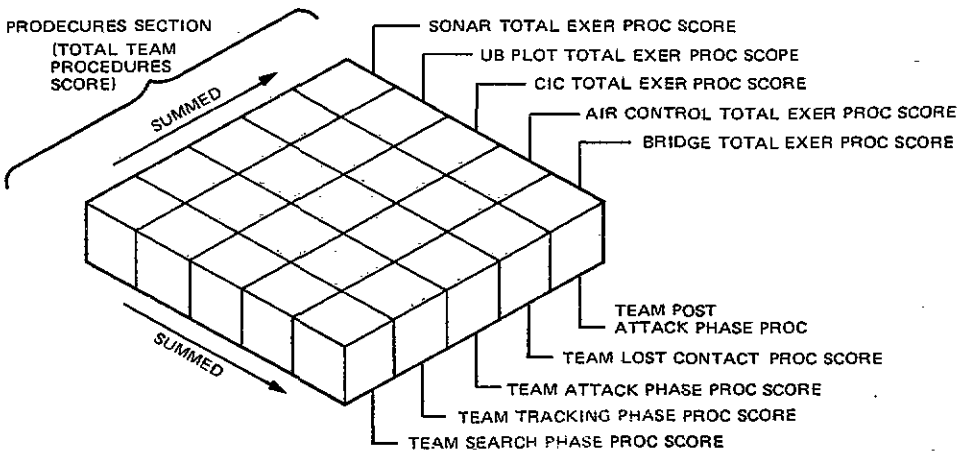

Figure 5. Measurement Category Scoring Composite

The detection of procedural errors by the computer can be accomplished if two requirements are met. These are:

1. The switch position or equipment operation must be sensed by the computer.

2. A standard problem, or portion of a standard problem, must be used so the doctrine procedure can be programmed for comparison.

In dynamic exercises where the team and computer interact and react, the number of doctrine switch settings which can be predicted is severly limited. The number of possible choices available to the operator must be limited by structuring the problem.

In these dynamic situations, an observer can be more effective. A well trained observer can monitor the operator's actions and record the error, time of occurrence and criticality. These equipment operator procedural errors can be combined with the other observer detected errors such as manual errors which are not available to the computer, i.e., manual plot entries. All such errors should be detected and included in the team performance evaluation. It should be noted here that the number of procedural errors will be more useful as a score than the number of correct actions. This count will then include both errors of omission and commission and will be much more useful in calling specific errors to the attention of the team for correction.

The need to weight errors in terms of significance is also a possibility. Weighting of scores in terms of criticality, or effect on mission success could affect the team score but would be of less importance in feedback to the team for correction.

When the procedural error score is within a satisfactory tolerance, training on procedure following can be considered as complete.

- Accuracy Measures

  Most measures of operator and team performance accuracy can be obtained by the computer. In fact, the computer is ideal for recording and making operator accuracy measurements. These measurements can include a great number of positioning, tracking and switch setting actions which are a measure of operator skill. Such measurements can be made periodically at one second, ten seconds or one minute intervals for continuous operations or when specific events occur. When continuous operations are involved, a beginning and an ending event is required to prevent making measurements at inappropriate times. For example, for sonar cursor positioning, accuracy should not be recorded before contact nor after time of fire.

  In addition to the computer recorded accuracy measurements, some accuracy measurement may need to be made by the instructor or observer. Manual plotting accuracy is one such example which is normally not available to the computer for measurement.

  The evaluation of accuracy measurements can also be made by the computer. The accuracy of a specific team can be compared by the computer with the accuracy of other teams, with a fleet criterion or with allowable tolerance bands. These comparisons are necessary for an evaluation of "how good" or "how bad." This information, as well as the raw accuracy data, should be supplied to the team as feedback and reinforcement. The actual measurement and when it occurred, can help the operator to improve performance. Graphic data can show trends, time of errors, variability of performance and bias. This specific data can be most beneficial for operator and team training and improvement.

- Communication Measures

  The current practice for team evaluation as shown by the ASW Escort Qualification Report, Figure 1, is heavily weighted with communications between various team members. Over 50% of the specific evaluation entries relate to oral reports or involve communication between team members.

  Unfortunately, computer interpretation of the spoken word is not yet feasible so observers must be used to sense and evaluate the communication between various members of the team.

  As in the case of procedure following, the detection and scoring of communication errors only is the preferred approach. This approach assumes that most communication will be appropriate, correct and will occur at the proper time. The observers can then concentrate on obvious errors which require correction. These errors can be classified under the following types and examples for team evaluation and scoring:

1. Errors in transmitted data, i.e., true bearing reported when relative bearing is correct.

2. Missed information report, i.e., target doppler not reported.

3. Improper communication procedure, i.e., non-doctrine choice of words or sequence of data.

4. Unnecessary communication, i.e., irrelevant comments or criticism during the problem.

If the observer can record and count these type of errors, a high count would obviously indicate the need for more training. Similarly, a low count would reflect the performance of a well trained team.

- Timing Measures

  The time required by a team to accomplish a specific result is a significant indication of team performance. A minimum time score usually indicates efficiency, effectiveness and absence of errors. Team scores can be a very useful summary indicator of the skill level of teams.

  Timing measures require a beginning event or signal and a terminating event or signal. These signals may be detected directly by the computer or may require a cue from the monitor. The computer records the time of the event and records the elapsed time as the measure. This time period can then be compared for team, individual, and problems.

  To be meaningful the team scores should only be used under standard problem conditions. When the initial problem conditions are identical, and carefully controlled, the variation in time between individuals or teams is due primarily to the skill level. It is assumed that the shorter the time required to reach the objective, the higher is the skill level of the team.

- Tactics Measures

  The measurement of the tactical moves made by a team in a given problem situation and the subsequent evaluation of these moves in terms of tactical effectiveness, adherence to doctrine and training, is in a different dimension than the previous measures. It is obvious that tactics is a significant factor in the overall performance of the team and in the evaluation of the team. However, the rationale and techniques which may be used for tactics measurement requires a more sophisticated approach than described above.

Our current effort, in conjunction with Decision Sciences, Inc., of San Diego, California, is the use of a Game Theory Approach which evaluates adversary tactical maneuvers. This effort is in the early stage of development and results are not available. However, it is anticipated that numerical values will become available. These numerical values will be converted into scores and combined with the team scores described above for a total team evaluation.

- Results to Date

  The ASW Tactics Team Trainers, located at Fleet ASW School at San Diego, California, have been programmed to provide computer printouts of problem data. The initial approach has been to print out all data in the data base once each second for analysis. The data for several pilot exercises were collected and analyzed in detail.

  Data for tactics evaluation and communication was not collected. Procedure following and timing data were available but meaningful evaluations could not be made. Accuracy measurements, however, proved to be very interesting. Figure 6 shows the sonar cursor bearing error from contact for a typical problem. Notice the magnitude of the error and the constant "lead" errors. Figure 7 shows sonar cursor range error for the same problems. Note the constant range error which shows a range lag on a closing target. Figure 8 shows a comparison between three runs on bearing error. Note the large variation between runs and also the constant "lead" error. Figure 9 shows similar data for comparative range error. Again the variability between runs is evident as well as the constant range error.

  The continuing effort will result in data on a larger number of teams as well as on a larger number of parameters. These data, together with observations by monitor personnel, will provide the necessary data for trainer evaluation and application of CAI.

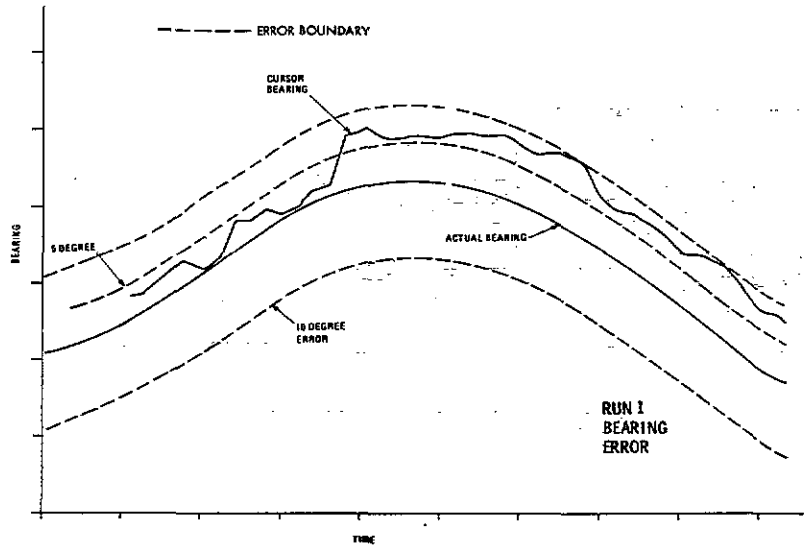

Figure 6. Run 1 Bearing Error

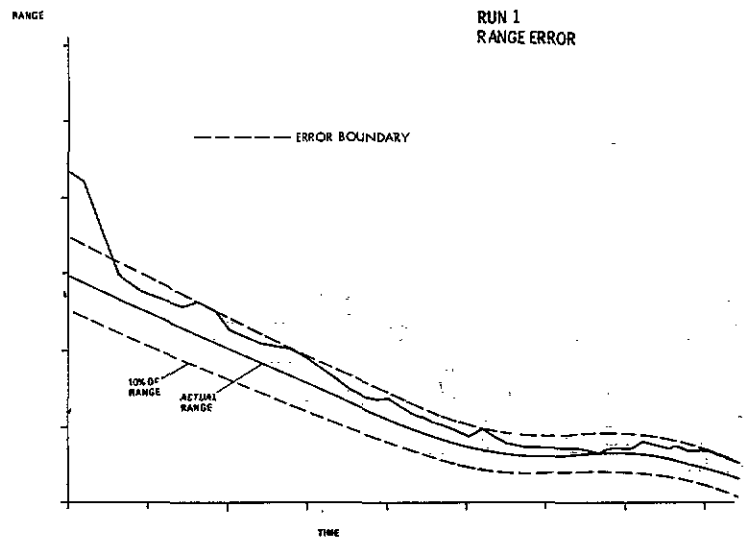

Figure 7. Run 1 Range Error

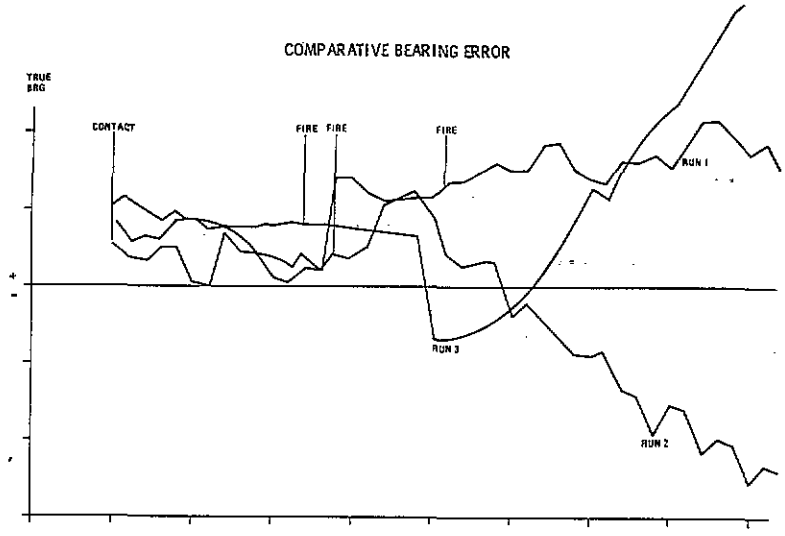

Figure 8. Comparative Bearing Error

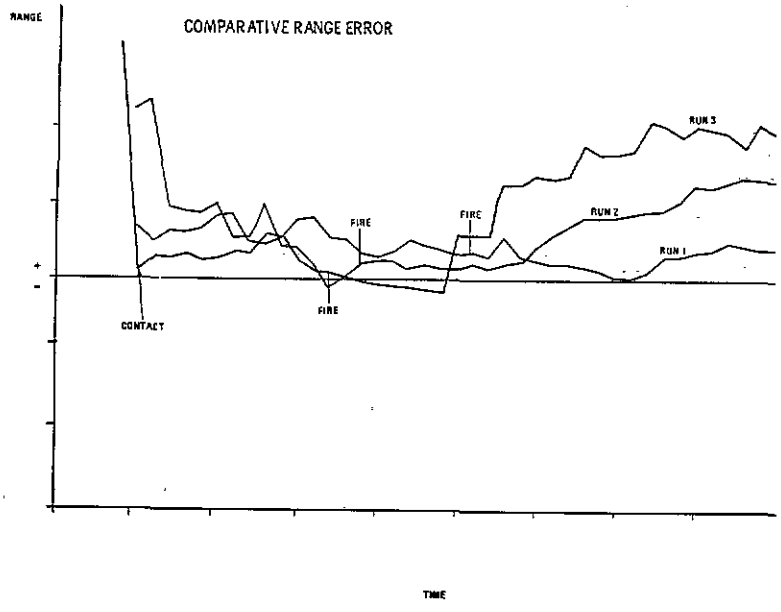

Figure 9. Comparative Range Error

References

a) Bitters, D. L. and Clark, G. M. Meaningful Composite Benefit Measures, Proceedings, Human Factors Society, 1975, October 1975.

b) Copperman, N. and Asa Dorian, P. Using CAI in Measuring Team Readiness, Proceedings, 9th NTEC/ Industry Conference, 1976.

c) Dickman, J. L., Automated Performance Measuring, Proceedings, 7th NTEC/Industry Conference, 1974.

d) Epstein, K. I. and Steinheiser, F. A. Bayesian Method for Evaluating Trainer Proficiency, Proceedings, 8th NTEC/Industry Conference, 1975.

e) Goldstein, I. L., Training: Program Development and Evaluation, Wadsworth Publishing, Belmont, California, 1974.

f) Shipley, B. D. (et al.), Simplifying the Measurement of Complex Skills in Training Simulator, Proceedings, 9th NTEC/Industry Conference, 1976.

g) Vruels, D. and Goldstein, I. In Pursuit of the Fateful Few: A Method for Developing Human Performance Measures for Training Control, Proceedings, 7th NTEC/ Industry Conference, 1974.

ABOUT THE AUTHORS

MR. BERNARD W. YAEGER is a Senior Principal Human Factors Engineer at Honeywell Marine Systems Center. For the past 20 years at Honeywell, he has performed Human Factors work on numerous Navy systems including underwater and surface sonar systems, ASW and Fire Control systems, and surface and subsurface warfare on individual and team trainers. His Human Factors' activity has covered the field from systems analysis and requirements, through design to final test and evaluation. Mr. Yaeger received his B. S. degree in psychology from the University of Iowa and his Masters degree in industrial psychology from Purdue University.

MR. JAMES D. BELL is a Human Factors Engineer at Honeywell Marine Systems Center. At Honeywell, he has worked on the development of ASW Team Performance Measurement models, Human Factors systems and analysis of ASW systems for the DD-963 Spruance Class Destroyer, and conducted personal performance evaluation and combat team evaluation for the DD-963. He was responsible for Human Factors engineering on the U.S. Air Force Undergraduate Navigator Training System (UNTS). Work on the UNTS project included development of test procedures for human engineering compliance, maintenance, and instructor and system effectiveness. Mr. Bell instructed the UNTS Instructor/Operator course for U. S. Air Force navigator instructors, and has performed general Human Factors systems analysis on other projects. He has worked on the systems crew efficiency analysis, and development of statistical analysis of crew performance on the P3 Orion ASW aircraft, while at Lockheed Aircraft Company. He received his B. A. degree in psychology from the University of California, Los Angeles; and his M. A. degree in psychology from California State University, Los Angeles.