PREDICTION OF TRAINING DEVICE EFFECTIVENESS FROM QUANTITATIVE TASK INDICES

A. Mirabella and G. Wheaton
Senior Research Scientists
American Institutes for Research

The problem of task fidelity is one which continually faces designers of training devices. Early during conceptualization of the device, decisions must be made concerning those features of the operational task which should be incorporated into the trainer in order to make the device optimally effective for both the acquisition and transfer of skills. Complementary decisions are needed concerning those features of the operational task which can be cost-effectively eliminated. Yet, objective means for making such decisions on an a priori basis (i.e., before the device is actually built) have never been developed. In particular, quantitative methods have been lacking with which to relate variations in trainer task characteristics to variations in skill acquisition and skill transfer. The pragmatic consequence of this lack of a predictive methodology has been incorporation into training devices--and, in particular, simulators--of as much realism as the state of the art and available dollars will permit. Increasingly, the cost effectiveness of such a response to training needs has been questioned.

A major stumbling block to the development of more objective and reliable approaches to device design has been the lack of a widely accepted and generic methodology for quantitatively analyzing and describing trainee tasks. It was the desire to work toward such a methodology which prompted NTEC to undertake a program of research on trainee task quantification and prediction of training effectiveness. Two major issues had to be addressed in the course of this program. First, would measures of training effectiveness (i.e., rate of skill acquisition, level of transfer) vary in some predictable manner as features of a training device were manipulated? Unless there was a relationship between these two sets of variables, prediction of effectiveness would not be feasible. Second, and even more basically, would it be possible to describe the critical features of a device reliably and along a number of quantitative dimensions? Unless such description were possible, there would be no way of investigating the relationship of interest.

To resolve these issues NTEC sponsored the American Institutes for Research in a program which had three objectives. These involved: (a) exploring ways of reliably describing features of trainee or instructor stations in quantitative terms; (b) determining whether the quantitative descriptors could be used to describe fairly complex devices; and (c) developing methods for relating the quantitative descriptors to variations in device effectiveness.

In pursuit of the first objective, a number of quantitative task-descriptive indices were assembled. These indices represented critical dimensions of the stimulus, response, and procedural aspects of trainee and instructor stations. Critical dimensions were those which, if manipulated, would be expected to affect level of (instructor) proficiency, rate of skill acquisition, or degree of transfer. Included among the indices were a variety of rating scales developed by AIR, and relating to such dimensions as work

load, precision of responses, and response rate. Other indices were based on metrics such as the Display Evaluation Index (DEI) developed by Siegel and his co-workers (Siegel, Miehle, & Federman, 1963) and the several panel layout metrics developed by Fowler and his associates (Fowler, Williams, Fowler, & Young, 1968). The DEI is basically a measure of the effectiveness with which information flows from displays, via the operator, to corresponding controls. The panel-layout indices represent the extent to which general human-engineering principles have been applied to the design of hardware.

To satisfy the second objective, the indices were applied to four sonarman trainee tasks (i.e., setup, detection, localization, and classification) as represented in a variety of different sonar training devices. This exercise demonstrated that most, if not all, of the indices could be used reliably to scale the extent and manner in which the trainee tasks differ across devices. By extension, therefore, the indices might be used to describe reliably and quantitatively how competing designs of the same device differ.

Many of the indices originally examined were excluded from the final battery because they could not be applied reliably or easily, or because they did not discriminate effectively among tasks. The 17 passing this hurdle were primarily the generic indices of Siegel, and also Fowler et al.

In response to the third objective of the program, that of validating the battery of indices which had been collated, a multiple regression model was employed which differed somewhat from the model conventionally used to predict individual performance from test scores. In the model which we employed, the predictor scores were task characteristic index values. The criterion was an average performance score for multiple subjects performing a single task. In other words, the model was designed to predict the average performance level on a particular task from index values descriptive of that task. This model has been applied in several different laboratory studies over the last two years.

These laboratory efforts have employed a modularized synthetic sonar trainer, constructed to represent a cross section of some 15 different sonar devices which had been previously task analyzed. The trainer consisted of 20 different modular panels representing different sonar console functions. For most of the functions there were alternatively designed panels which could be interchanged, and thus used to manipulate the overall design of the trainer console. Figure 1 shows a photograph of one such console configuration. This was defined as our most complex configuration. Note, for example, the panel at the top left. This panel represents the function of energizing the console. It consists of a number of toggle switches, feedback lights, a rotary switch, and a meter. In other configurations of the console, this particular panel might be replaced by one which consists of nothing more than one toggle switch and one feedback light. Similarly, most of the other panels were designed in alternative forms: a "simple" version and a "complex" version for accomplishing basically the same function.

Through appropriate use of panels, there were a number of ways in which the operator's task could be manipulated. For instance: 1) alternative panels could be employed; 2) one task could be embedded in a second task by

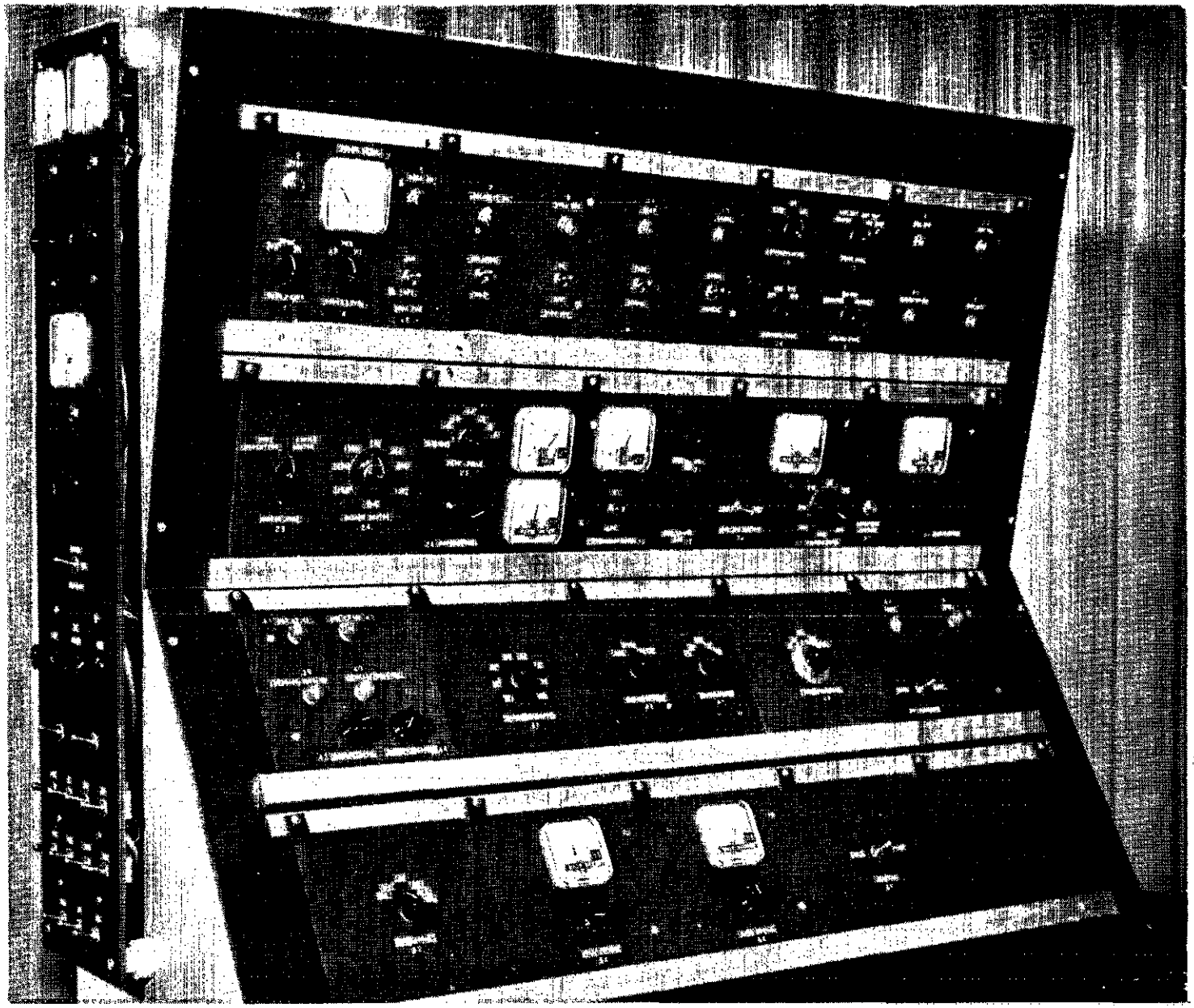

Figure 1. Modularized Sonar Trainer, Complex Version.

making some of the displays and controls irrelevant for performance of the first task; 3) feedback lights associated with toggle switches could be masked; and 4) contingency responses could be built into the training procedure. These various manipulations were employed and then the task characteristic battery was applied to the results of the manipulations. Twenty different tasks were generated in this manner, for each of which there was a corresponding set of task index values. These tasks provided the basis for the laboratory validation experiments.

The first of these experiments focused upon acquisition of "sonar" set-up skills. Could acquisition of skill be related to variations in the 17-index battery; and if so, how consistent would these relationships be across a series of training trials?

In this first laboratory study, performance data were collected for 12 tasks out of the 20 which had originally been generated. These 12 tasks were chosen because they were widely dispersed in terms of the task characteristic indices. For each of these tasks, criterion performance data were provided by average group scores.

The general method of instruction employed in this and subsequent laboratory experiments was to describe to the trainee the entire setup procedure, twice in succession. Each response in the procedure was indicated to the trainee, along with a verbal statement which he was to make as he performed a particular operation. For example, he might have been told to set the power switch, #1, to the "on" position, and say "#1 to 'on'." Verbalization by the trainee was necessary to facilitate the recording of incorrect or omitted responses in the subsequent test trials. The test technician could identify these errors by following a procedural check list, and noting where the trainee deviated from expected verbal statements. A stopwatch record of total performance time for each test trial was maintained.

Following the initial two orientation trials, the trainee was exposed to 15 test trials, each involving a complete run-through of the setup procedure for that particular task. He was corrected for any wrong or omitted responses--the stopwatch being halted while corrective instructions were given. It should be emphasized that following each trial the settings of all controls were scrambled so that the initial appearance of the console varied somewhat from trial to trial. Furthermore, there were a number of specific instructions which changed from trial to trial. As an example, the trainee might have been instructed to set up for passive sonar search on one trial, and for active sonar search on the subsequent trial. The specific sequence of required responses varied accordingly. The point which is to be made here is that we tried to make these tasks logical rather than merely rote activities.

The results for the acquisition study were very encouraging. First, data plots for the various tasks and across testing time showed substantial variation in both performance times and performance errors. This capacity to differentiate among the tasks in terms of the criteria was essential if we hoped to have any success in generating significant multiple correlation coefficients.

Next, rather substantial multiple correlations were obtained between the task indices and both time and error scores. But before discussing these any further, it should be pointed out that the criterion data were initially corrected for the effects of task length through the use of partial correlation. Task length had a very substantial impact upon performance, but was of trivial interest in this experiment. The full set of 17 predictors was also reduced to three predictors through the use of a step-wise regression procedure, a technique which essentially identifies an optimum subset of predictors.

The results of the regression analysis are shown in Table 1.

TABLE 1. · SUMMARY OF MULTIPLE REGRESSION ANALYSES OF PERFORMANCE TIME AND NUMBER OF ERRORS FOR FIRST, MIDDLE, AND LAST BLOCK OF ACQUISITION TRIALS

Criterion	R	R^2	df	F	Indices in order of selection by step-wise regression program
			Time Scores		
T_{1-2}	.780	.608	3, 8	4.69*	E%, AA%, D%
T_{7-8}	.744	.553	3, 8	3.30	E%, AA%, DISP
T_{13-15}	.626	.392	3, 8	1.72	AA%, C%, DISP
			Error Scores		
T_{1-2}	.651	.423	3, 8	1.96	E%, C%, D%
T_{7-8}	.896	.802	3, 8	10.80**	AA%, MAIN, D%
T_{13-15}	.875	.766	3, 8	8.73**	AA%, CONT, DEI

*$p < .05$.
**$p < .01$.

This table summarizes correlational data for both time scores and error scores at three points in the testing--early, middle, and late. For example, T_{1-2} refers to testing trials 1 and 2 combined. T_{7-8} refers to the middle trials 7 and 8 combined. T_{13-15} refers to the last three testing trials combined.

For each time block, the table provides a multiple correlation coefficient (R), the percentage of variance in the criterion accounted for by the selected task indices (R^2), degrees of freedom, and the corresponding F

statistic. Finally, the indices accounting for the correlation are provided. Appendix A defines what each of these indices is.

Note, first, that while both errors and time are significantly related to subsets of task indices, they are not related at the same points in testing. Performance time is related to the indices early in testing, specifically at T_{1-2}, while performance errors exhibit their covariation with the indices later in testing, specifically at T_{7-8} and T_{13-15}. This is very reasonable. It suggests that while the trainee is still becoming familiar with the task, he adjusts his speed to match the difficulty of the task. In this way he minimizes errors. Later in testing, when familiarity has set in, this adjustment is no longer made and error rate is more closely tied to task complexity.

The second point worth noting is that the specific patterns of salient indices change from time scores to error scores. In fact, the pattern varies across testing periods. There is some overlap in the predictors, but this overlap is far from perfect. To the extent that these variations are meaningful and not merely due to unstable data, they point up the criterion problem which may underlie assessment of device effectiveness. These variations do not provide great comfort for a simple figure of merit approach to prediction of device effectiveness, at least insofar as skill acquisition is concerned.

Having demonstrated that quantitative task indices could be related to the acquisition of procedural task skill, attention turned to the issue of transfer of training. Could those same indices predict transfer? The first experiment was essentially repeated on new subject samples with the following additional task. All subjects, subsequent to the initial 15 acquisition trials, were trained and tested for 10 trials on a common task of medium complexity. Thus, some groups of subjects transferred from difficult tasks to the intermediate task, while others transferred from relatively easy tasks to the intermediate task. Comparisons of transfer performance were based upon the common intermediate task. The criteria of interest were the actual time and error scores achieved on the second or transfer task. For this experiment the predictor scores were transformed into difference scores, as a way of indexing fidelity of simulation. That is, we computed absolute differences between task index values for the transfer task and values for the initial training tasks.

The results of the transfer study are shown in Table 2. A number of interesting contrasts are seen here vis-à-vis the acquisition data described in Table 1. As before, significant relationships are found with task indices, but the relationships are considerably more complete and uniform. Significant and rather substantial correlations appear at each point in testing, both for error and for time scores. Furthermore, there is a gratifying consistency of predictors within each type of criterion. Consistency is perfect for the error scores and very high for the time scores. Thus, we have more encouragement here for a figure of merit approach. We still have to reckon with a change in the pattern of predictors as we go from time to error criteria, but at least within either criterion there is consistency.

TABLE 2. SUMMARY OF MULTIPLE REGRESSION ANALYSES OF PERFORMANCE TIME AND NUMBER OF ERRORS FOR FIRST, MIDDLE, AND LAST BLOCK OF TRANSFER TRIALS

Criterion	R	R^2	df	F	Indices in order of selection by step-wise regression program†
			Time Scores		
T_{1-2}	.72	.51	3, 11	3.88*	DISP, C%, FBR
T_{5-6}	.76	.58	4, 10	3.50*	DISP, C%, DEI, AA%
T_{9-10}	.84	.71	4, 10	6.16**	DISP, C%, D%, FBR
			Error Scores		
T_{1-2}	.78	.61	4, 10	3.93*	DEI, E, DISP, CONT
T_{5-6}	.90	.81	4, 10	10.33**	DEI, E, DISP, CONT
T_{9-10}	.84	.71	4, 10	5.98*	DEI, E, DISP, CONT

†Index values represent difference scores.
*$p < .05$.
**$p < .01$.

To summarize, it has been possible to demonstrate with this series of experiments that variations in quantitative task indices can be related significantly and consistently to trainee performance, at least where transfer criteria were employed (Wheaton, Mirabella, & Farina, 1971; Wheaton & Mirabella, 1972a, 1972b). Given this modest success, the step remains to be taken of essentially repeating the laboratory work in a field environment, using sonar trainers or, preferably, a synthetic trainer and related operational stacks.

We should also like to stress that while the focus of the research just described was upon trainee task variables, it is recognized that this class of variables is not the only one which impacts upon device effectiveness. Training method, including device utilization, may be as potent, if not more so. In the final analysis, task variables, instructional variables, and individual differences have to be built into the "effectiveness equation," particularly because these may interact in critical ways.

One example of such an interaction is provided by an experiment which was conducted toward the end of our task quantification project. This experiment

represented a variation on a previously studied theme. A comparison was made among hot panel, cold panel, and photographic representations of the modularized trainer which had been used in the earlier phases of research. The variation was a manipulation of task complexity as defined by our quantitative indices. Complex, intermediate, and simple versions of the trainer were employed. The results were interesting in view of the usual conclusion--namely, that the hot panel is not critical for procedural training. A clear-cut interaction was found between task complexity and mode of console presentation. Variance associated with presentation mode narrowed systematically as task complexity decreased. The hot panel was distinctly superior for the complex task, but lost its advantage for the simple task. Results of this sort underscore the importance of being able to quantify task variables and, thus, being able to pigeonhole tasks or training devices "by the numbers." The program of research which has been summarized here represents, we hope, some progress in this direction.

REFERENCES

Fowler, R. L., Williams, W. E., Fowler, M. G., & Young, D. D. An investigation of the relationship between operator performance and operator panel layout for continuous tasks. AMRL-TR-68-170 (December, 1968). Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, Ohio.

Siegel, A. I., Miehle, W., & Federman, P. Information transfer in display-control systems. VII. Short computational methods for and validity of the DEI technique (Seventh Quarterly Report, 1963). Applied Psychological Services, Wayne, Pennsylvania.

Wheaton, G. R., & Mirabella, A. Quantitative task analysis and the prediction of training device effectiveness. Proceedings of the 5th Annual NTDC/Industry Training Conference, February, 1972a.

Wheaton, G. R., & Mirabella, A. Effects of task index variations on training effectiveness criteria. Technical Report No. NAVTRADEVCEN 71-C-0059-1, 1972b. Naval Training Device Center, Orlando, Florida.

Wheaton, G. R., Mirabella, A., & Farina, A. J., Jr. Trainee and instructor task quantification: Development of quantitative indices and a predictive methodology. Technical Report No. NAVTRADEVCEN 69-C-0278-1, 1971. Naval Training Device Center, Orlando, Florida.

APPENDIX A

TASK CHARACTERISTIC INDICES

1. MAIN - defined as the number of responses comprising the main or dominant procedural sequence in an operations flow chart.

2. CNTG - defined as the number of responses comprising the auxiliary or contingency procedural sequences.

3. TA - defined as the total number of responses (actions) comprising the procedural sequence in an operations flow chart. It represents the sum of MAIN and CNTG.

4. CONT - defined as the total number of different controls manipulated during performance of a subtask.

5. DISP - defined as the total number of different displays referenced during performance of a subtask.

6. E - defined as the total number of different equipment elements interacted with, this index is given by the sum of CONT and DISP.

7. LV - the link value reflecting the relative strength of the sequence of use among the various controls and displays. As used here, it is the sum of the products of the number of times a link is used, and the percentage of use of the link (Fowler, Williams, Fowler, & Young, 1968).

8. AA% - an index reflecting the percentage of alternative actions present in an operation. A score of "0% means that the highest number of alternative links are used, each with an equal frequency of use, and 100% score means there is only one link out of and into each control, with the same frequency used for all links" (Fowler et al., 1968).

9. F% - another index (Fowler et al., 1968) describing the extent to which all controls and displays are used an equal number of times (0%) or a theoretically defined optimum number of times (100%).

10. DEI - a measure of the effectiveness with which information flows from displays via the operator to corresponding controls. The index yields a dimensionless number representing a figure of merit for the total configuration of displays and controls (Siegel, Miehle, & Federman, 1962).

11-13. D%, C%, E% - defined respectively as the number of display, control, or combined equipment elements which the operator actually employs relative to the total number of such elements which are available for use.

14-17. CRPS, FBR, INFO, INST - refer to the frequency with which the operator makes various types of responses during performance of the task. Included are responses involving manipulation of controls (CRPS), securing of feedback (FBR), acquisition of information (INFO), as well as those primarily initiated by the instructor (INST).

ABOUT THE AUTHORS

DR. ANGELO MIRABELLA, Senior Research Scientist, received his A.B. degree from Cornell University in 1960 and his M.A. from Columbia University in 1961. In 1964, he received a PhD from the University of Massachusetts in Experimental Psychology.

From 1964 to 1969, Dr. Mirabella was employed as a Research Scientist at the Electric Boat Company, Groton, Connecticut. While there he directed and conducted research in a number of areas related to command/control displays for submersible systems. Areas included adaptive training of monitoring skills, effects of ambient noise on sonar monitoring, effects of display format on operator control of deep submergence search vehicles, and effects of digital display format on initial target detection in an ASW context. The results of these efforts were reported in a series of in-house, government, and professional publications.

Particularly relevant was his work for the Naval Training Equipment Center. This effort explored the use of computer-controlled displays for adaptive training of visual monitoring skills.

His experience in subsystem analysis has included the following: (1) task analysis of a man-computer target ranging subsystem within an ASW fire control system, and (2) analysis of monitoring performance for a projected 30-hour submerged search operation, employing side-looking sonar, under a variety of proposed crew configurations and corresponding work/rest cycles. This latter effort provided basic performance data for operations analytic modeling of a deep submergence search vessel.

As a consultant to the Personnel Department at Electric Boat, he also performed job analyses. His efforts were summarized in several in-house reports, including a Handbook of Skills for the Job of Electrician.

Since joining AIR in 1970, he has been responsible for development of a program of research dealing with the psychophysiological effects of noise pollution. As part of this responsibility, he supported a program of research for NASA aimed at developing a standardized performance battery for use in assessing the effects of environmental stressors.

Dr. Mirabella has also been involved in a number of human engineering projects. Since joining AIR, he has provided project support on an NAVTRAEQUIPCEN contract for trainee and instructor task quantification.

Dr. Mirabella was principal investigator for a recently concluded ONR-sponsored project to revise the Human Engineering Guide to Equipment Design.

He is currently Co-Principal Investigator for a third year continuation of AIR's task quantification project, sponsored by NAVTRAEQUIPCEN. The purpose of this effort is to validate a number of task analytic indices, using transfer of training measures as criteria both in the laboratory and under field training conditions. An additional aim is to measure interactions between task characteristics and training methods, using a sonar simulator as the research vehicle.

Dr. Mirabella is also currently Principal Investigator for a project sponsored by the Secretary's Commission on Medical Malpractice. He has been responsible for conducting survey research on characteristics of patients, health care workers and incidents involved in medical injury, using insurance industry files as the data source.

MR. GEORGE R. WHEATON, Senior Research Scientist, received his B.A. degree in Psychology from Bowdoin College in 1961 and his M.S. in Experimental Psychopathology from McGill University in 1963. He has taken additional graduate level work at the George Washington University.

In May of 1963, Mr. Wheaton joined the Missile and Surface Radar Division of the Radio Corporation of America as a Personnel Specialist. In this capacity he was responsible for developing selection test batteries for a variety of personnel positions and assisted in the design and implementation of personnel training programs. He also participated in an innovative research program concerned with modification of managers' attitudes and behaviors as a function of sensitivity training by the T-group method.

Mr. Wheaton subsequently joined the Bio-Sciences Program at RCA's Advanced Data Systems Center. His work at the Center involved both basic and applied research on a variety of human factors problems associated with a broad range of man-machine systems.

Since joining AIR in 1966, Mr. Wheaton has served in a senior technical capacity on a number of projects of potential relevance to the proposed research. He has participated in a variety of drug studies conducted for both military and civilian organizations. The emphasis in these studies has been in relating obtained effects to a variety of individual difference variables. In one study of relevance, a variety of tests was used to compare subjects classified as eligible or ineligible for testing under certain types of drug treatments. A variety of other projects have given him experience in both individual and group testing under field and classroom conditions.

More recently, Mr. Wheaton has directed a series of projects for the Naval Training Equipment Center. This long-range research program has focused on quantification of design features of complex training devices and on the relationship between such parameters and measures of training effectiveness.