Special Report 26

Principal Scientist Colloquium

May 1996

United States Army Research Institute
for the Behavioral and Social Sciences

REPORT DOCUMENTATION PAGE

1. REPORT DATE (dd-mm-yy) May 1996	2. REPORT TYPE Final	3. DATES COVERED (from. . . to) June 1993-June 1995

4. TITLE AND SUBTITLE Principal Scientist Colloquium	5a. CONTRACT OR GRANT NUMBER 622785
	5b. PROGRAM ELEMENT NUMBER A791
6. AUTHOR(S) (in alphabetical order) U.S. Army Research Institute	5c. PROJECT NUMBER 8005
	5d. TASK NUMBER C01
	5e. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) U.S. Army Research Institute for the Behavioral and Social Sciences 5001 Eisenhower Avenue Alexandria, VA 22333-5600	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Research Institute for the Behavioral and Social Sciences 5001 Eisenhower Avenue Alexandria, VA 22333-5600	10. MONITOR ACRONYM ARI
	11. MONITOR REPORT NUMBER Special Report 26

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT *(Maximum 200 words)*:

In June of 1995, the U.S. Army Research Institute for the Behavioral and Social Sciences held a Principal Scientist Colloquium in which fifteen scientists from nine ARI research units presented results from their recent in-house research. This report presents the top three papers from the colloquium plus abstracts from all the other papers. The three top papers, printed in their entirety, are: (1) Unaided Night Vision Training; (2) How to Make Decisions About the Effectiveness of Device-Based Training; and (3) Understanding and Improving Tactical Problem Solving.

15. SUBJECT TERMS

Night vision training Device-based training Tactical problem solving

SECURITY CLASSIFICATION OF					
16. REPORT Unclassified	17. ABSTRACT Unclassified	18. THIS PAGE Unclassified	19. LIMITATION OF ABSTRACT Unlimited	20. NUMBER OF PAGES 66	21. RESPONSIBLE PERSON (Name and Telephone Number) David Witter 703-617-0324

Principal Scientist Colloquium

May 1996

United States Army Research Institute
for the Behavioral and Social Sciences

FOREWORD

The U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) is the Army's lead laboratory in developing and fielding personnel performance and training technology. With roots reaching down to the first soldier selection efforts in 1917, ARI provides research on recruitment, selection and assignment, retention, and force management to assure quality, trained soldiers and units for America's Army. ARI's research seeks to understand the underlying skills, knowledge, and experiences that are important for acquiring, training, developing, and deploying soldiers, as well as to develop new methods for using emerging technologies to improve personnel performance and training.

In 1992, ARI created a dual track career program in which senior research professionals are given career opportunities as either full-time research managers or full-time researchers/principal scientists. The program recognizes the distinct value of both the management and technical tracks, providing opportunities and resources for challenging scientific and technical work, as well as for developing effective research management and administrative skills. This report presents recent products from ARI researchers in the principal scientist track.

On 22 June 1995, ARI held a Principal Scientist Colloquium in which 15 top scientists from 9 ARI research units presented results of their recent in-house research. This ARI Special Report presents the top three papers from the colloquium along with abstracts of all of the other papers. The top papers were selected by a panel of judges that included Dr. Steve Sellman, Director of Accession Policy, Office of the Secretary of Defense, and Dr. Jesse Orlansky, Institute for Defense Analysis. The panel evaluated the researchers and their work on three factors: (1) quality of science, (2) productivity, and (3) relevance to the Army.

Dr. Jean Dyer from ARI's Infantry Forces Research Unit received the Hubert E. Brogden Award for Research Excellence for her paper, *Unaided Night Vision Training*. Her research developed, refined, and evaluated a training program that improves soldier performance during night operations by maximizing their effective use of unaided vision. Dr. John Boldovici from ARI's Simulator Systems Research Unit and Dr. Jon Fallesen from ARI's Fort Leavenworth Research Unit received Honorable Mentions for their work. Each of these papers is included in this report.

The first ARI Principal Scientist Colloquium was a great success. The presentations were enlightening, interesting, and demonstrated the high quality of ARI's in-house research and its researchers. The research presented at this colloquium and summarized in these abstracts represents the products of more than $1.5M investment in in-house research by ARI and the Army for solving current and future manpower, personnel, and training (MPT) needs. The results of this and other ARI research provide America's Army with the MPT technologies needed for success in military operations such as peacekeeping and for success on the battlefield.

EDGAR M. JOHNSON
Director, ARI and
Chief Psychologist,
U.S. Army

Table of Contents

Unaided Night Vision Training

Jean L. Dyer, Ph. D.
U.S. Army Research Institute
Infantry Forces Research Unit, Fort Benning

The extent to which soldiers and units depend upon their vision to conduct tasks and missions effectively becomes very obvious during night operations. Many technological improvements in the equipment used by soldiers have been generated because of the need to improve the ability to see the battlefield at night. Nevertheless, regardless of the technologies available, all soldiers must still understand and master the basics of unaided night vision, how to maximize their night vision capabilities, and to be confident in operating at night with only their eyes.

The NIGHTFIGHTER Research Program

This paper describes several experiments conducted from 1993 through 1995 on an unaided night vision training program developed specifically for ground forces. The research was part of the NIGHTFIGHTER research program at the Infantry Forces Research Unit (IFRU). The goals of NIGHTFIGHTER are to identify the critical and most frequent problems encountered during night operations, followed by the identification and testing of possible training solutions. In addition to the unaided night vision research reported here, a front-end analysis of critical night operations problems was conducted, and solutions were found to the problems in zeroing aiming lights to the M16A2 rifle when firing with night vision goggles. Current research is on determining the best field-expedient techniques for adjusting the acuity of night vision goggles, training programs for thermal target acquisition and identification, and improving leaders' ability to train small units for night operations, specifically the night attack.

The front-end analysis revealed little research on night operations beyond that of operational tests of night equipment. However, training was always identified by soldiers and leaders as vital to night performance. Analysis of the training and doctrine literature, training materials, and research on night operations showed failures to communicate to soldiers what was already known about night operations and failures to investigate critical ground force training issues. The NIGHTFIGHTER research program focuses on both these training deficiencies.

Need for Training Program on Unaided Night Vision

Knowledge of how the eye functions at night, that is, the psychophysics of the eye, is not new. Much basic research was conducted in the late 1930s and early 1940s. As shown in our review of Army field manuals and publications in the military literature, unaided night vision training for ground forces existed as early as World War II.

However, most of what is known about the eye was not in the ground forces' literature and training when the current unaided night vision training research began. In addition, some errors and misconceptions about night vision were in print. When interviewed, many soldiers indicated they had inadequate training. Responses ranged from privates who said the only thing they had been taught was to close one eye when firing a rifle, to senior noncommissioned officers who had instruction in basic training, to other senior leaders who said they had not heard about unaided night vision since Vietnam. Clearly, here was an instance of where we had failed to communicate to soldiers what was known about how the eye functions at night and how to maximize their night vision.

The Unaided Night Vision Program

The unaided program is 45 minutes long and is presented entirely in the dark via 35-mm slides. Neutral density filters are used on the slides to control the intensity of the light and to allow individuals to dark adapt over the instructional period. The application of this technology to instruction on night vision was developed by Cdrs M. H. Mittelman and D. L. Still of the Naval Aerospace Medical Research

Laboratory and is currently patent pending. The slides present basic information on unaided night vision. However, the unique feature of the program is the demonstrations which are provided via the specially constructed slides. While the eyes gradually adapt to the dark, demonstrations show what happens to vision at night and techniques to reduce visual illusions and other problems encountered at night.

The major program demonstrations are as follows:

- Time to Dark Adapt: Illustrated by the contrast between the first and last slides which depict three attacking soldiers. This scene is not visible at first, but is clearly visible at the end of the program.

- Night Blind Spot and Diamond Viewing Technique: Several times throughout the program soldiers stare at objects or small lights, and they disappear. Diamond viewing is shown as a way to maintain these objects in their field of view.

- Reduced Visual Acuity: Most of the word slides simulate the acuity typical of twilight, that is, 20/50. A silhouette scene with trees, buildings, radio towers, and telephone poles illustrates the inability to discriminate details of objects at night.

- Perception of Color and the Purkinje Shift: Two slides contrast how colors are perceived as shades of gray under low illumination and as distinct colors under high illumination. Red and green dots of light are presented to show how red loses its intensity and may fade away when viewed with peripheral vision, while green gets brighter. The demonstration also shows how both colors may disappear and only white light is seen.

- Autokinetic Illusion: A single source of light is presented; it appears to move as soldiers look at it. They are then shown how to reduce the apparent movement of the light by scanning.

- Effects of Lights on Dark Adaptation: The effect of short bursts of light such as that from strobes and tracers is shown. The effect of looking directly at a flood light on the dark adaptation of both a covered and uncovered eye is shown.

The structure of the ground forces' program was based on the Navy's aviator program (Mittelman & Still, 1989). Considerable changes and additions were made to tailor the program to the ground force audience. Demonstrations directed specifically to aviators were removed and ground demonstrations were added. Ground examples of night vision problems were added, based on input from soldiers with extensive night operations experience. The instructional guide was modified to include the instructional purpose of each slide, the concepts and examples to be stressed by the instructor, detailed instructions on how to give each demonstration, and a suggested script for each slide. Throughout program development, the Navy's expertise was used to ensure the scientific accuracy of the material and to produce the program. As there had been no evaluation of the aviation program, the experiments reported here constitute the only assessment of the effectiveness of the training media and the program content.

Evaluation With Experienced Soldiers

The ground forces' program was evaluated with soldiers who had different years of Army service as well as with civilian and military instructors (Dyer, Gaillard, McClure, & Osborne, 1995). Two experiments were conducted using the design presented in Table 1. The extent to which the program increased soldiers' knowledge of unaided night vision (posttest scores for the Program group) beyond current training and experience (initial test scores for the No Program group) was examined. In order to allow everyone to receive the program, soldiers in the No Program group were given the program after taking their initial test and were then retested (the follow-on activity column in Table 1).

Table 1
Experimental Design for Program Evaluation with Experienced Soldiers

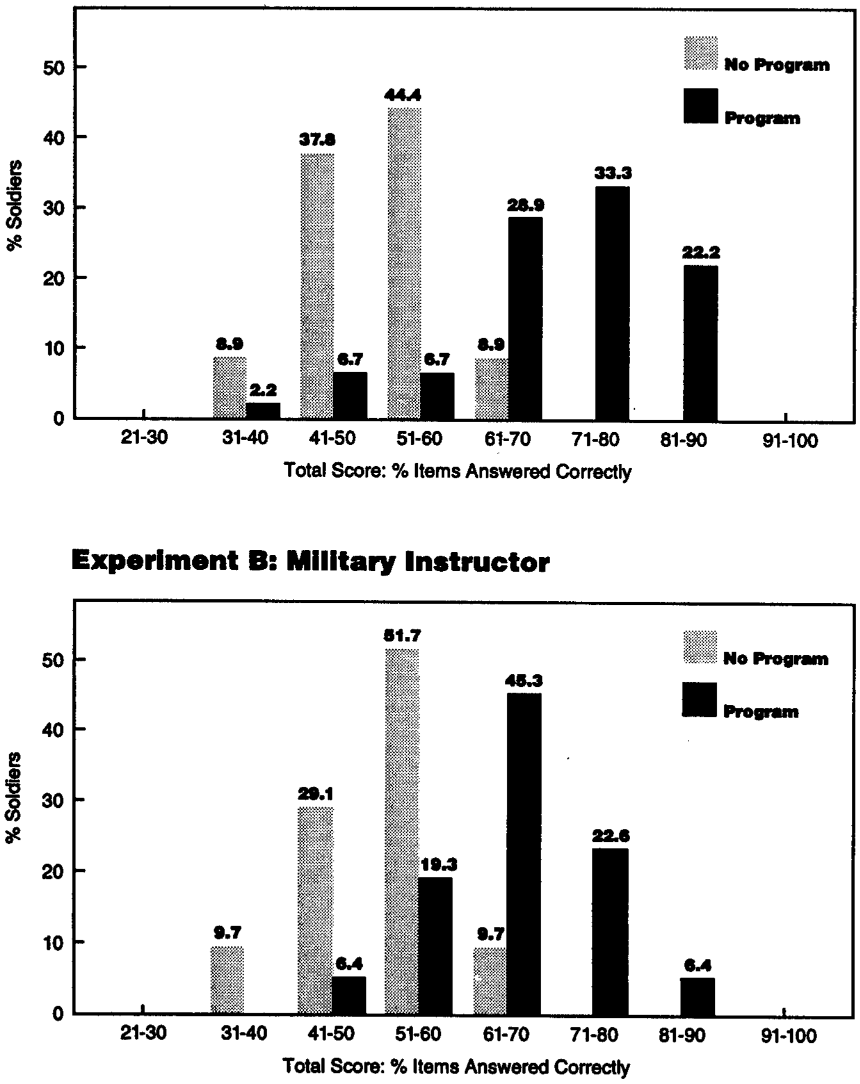
Experimental Conditions			Follow-On Activity	
R	Unaided Program Group Exp A: n = 45 Exp B: n = 31	Posttest		
	No Program Group Exp A: n = 45 Exp B: n = 31	Initial Test	Administration of unaided program to the No Program Group	Retest

Note. R stands for random assignment to experimental conditions.

Figure 1.

Distribution of scores for the Program group on the posttest and the No Program group on the initial test in experiments with experienced soldiers.

Experiment A: Civilian Instructor



Experiment B: Military Instructor

The difference in the two experiments was that a civilian instructor was used in one (Experiment A), and a military instructor was used in the other (Experiment B). Consequently, the experiments also allowed an examination of program effects as a function of instructor familiarity with the program content. The civilian instructor, a member of the research staff, was very familiar with the content, whereas the military instructor was not as familiar. Inclusion of a military instructor provided an assessment of the potential of program success in typical Army settings as well.

The posttest, initial test, and retest were identical, a 50-item test on unaided night vision. Subscores on important and less important content were obtained. In addition, three subscores on how the information was presented in the program were derived: information related to the demonstrations, more technical information which was typically presented on word slides, and items which required application of night vision concepts to new situations.

Experiment A was replicated three times; Experiment B, twice. Table 2 shows the years of Army experience for the soldiers in the two experiments. Time in service ranged from an average of 2.7 years to an average of 10.9 years.

Table 2
Soldier Experience: Mean Years in Army

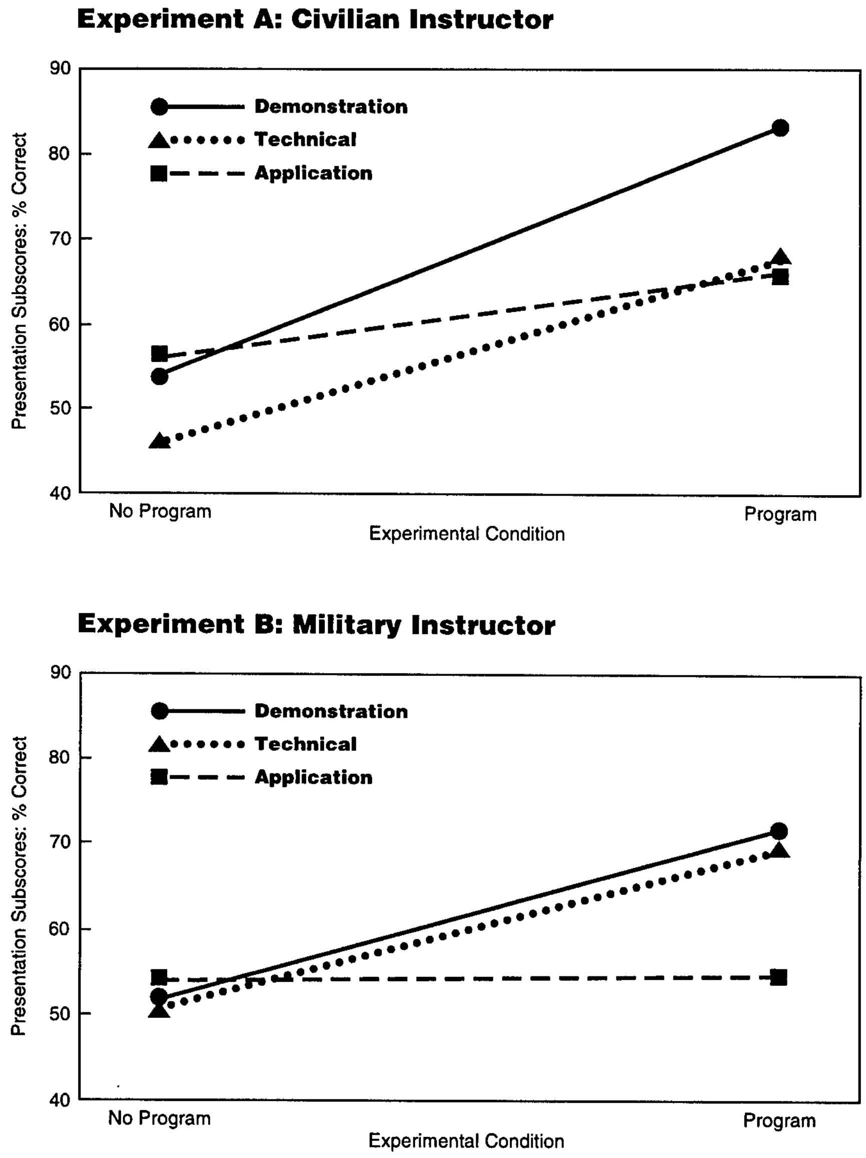
Soldier Category	Experiment A	Experiment B
Small-unit Leaders, FORSCOM Unit	5.9	6.8
Instructors & Cadre	6.0	
Ranger School Students	2.7	
Active & Reserve Component Leaders		10.9

Note. n = 30 per cell, except for the active and reserve component leaders where the n = 32.

The program was found to significantly increase soldiers' test scores, regardless of the length of time in the Army and whether the instructor was military or civilian (see Table 3). Posttest scores for the Program group were 1.3 to 1.4 times higher than the initial test scores for the No Program group. Figure 1 shows the frequency distributions for the two groups. The important material was acquired better than the less

Figure 2.

Presentation subscores for the Program group on the posttest and the No Program group on the initial test in experiments with experienced soldiers [$F(2, 168) = 16.73$, $p < .0001$].



Experiment A: Civilian Instructor

Experiment B: Military Instructor

important material, $F(1, 84) = 81.25, p < .0001$. The program also had the strongest impact on soldiers' knowledge of demonstration-related material and technical information. There was, however, less impact upon soldier ability to apply unaided night vision principles and concepts (see Figure 2).

Evaluation With Infantry Trainees

With Infantry trainees, we (Dyer et al., 1995) compared the knowledge gained from the program to that from reading the same material, with no exposure to the visual demonstrations. This was analogous to comparing what would be gained from reading a field manual with information on night vision to receiving a lecture-demonstration of the same concepts. The Infantry one-station-unit-training (OSUT) trainees in this experiment had not yet started their basic training, and therefore had no previous training in or experience with military night operations. At least 60% had a high school education; the mean age was 20.5 years; the mean General Technical (GT) score on the Armed Services Vocational Aptitude Battery (ASVAB) was 107. A separate Baseline group of 30 trainees was included. They had no exposure to the unaided program and were given the retention test. Their age, GT scores, and high school education backgrounds were the same as the other trainees.

Table 3
Results With Experienced Soldiers: Mean Percent Correct on Total Score

Soldier Category	Group		
	Program Posttest	No Program Initial Test	No Program Retest
Experiment A: Civilian Instructor			
Small-Unit Leaders	71	50	70
Ranger Students	71	50	76
Instructors & Cadre	69	53	74
All Soldiers	71	51	73
Experiment B: Military Instructor			
Small-Unit Leaders	63	51	72
Active & Reserve	68	53	72
All Soldiers	66	52	72
Both Experiments			
All Soldiers	68	51	73

Note. Program-No Program comparisons on posttest and initial test.
Experiment A: $F(1, 84) = 85.25, p < .0001$
Experiment B: $F(1, 58) = 40.60, p < .001$

The experimental design is in Table 4. A primary comparison in this experiment was between trainees who received the demonstration version of the program, the Program condition, and those who read a written version of the script, the Text condition. A retention test was given to both groups 24 days later. Program and Text condition scores were also compared to the Baseline condition.

The two versions of the program were equally effective overall, on the immediate posttest as well as on the retention test (see Table 5). Posttest scores were 1.5 times higher than Baseline scores obtained from trainees who were given no instruction on unaided night vision. Retention of the material remained high for both experimental conditions, as scores dropped little, being 1.3 times higher than the Baseline group.

Table 4
Experimental Design for Program Evaluation With Infantry Trainees

Experimental Conditions			Follow-On Activity	
R	Unaided Program (Demonstrations)	Posttest n = 41	24 Days	Retention Test n = 30
	Text Version	Posttest n = 41		Retention Test n = 35
	Baseline Group, n = 30. No instruction on unaided night vision.			

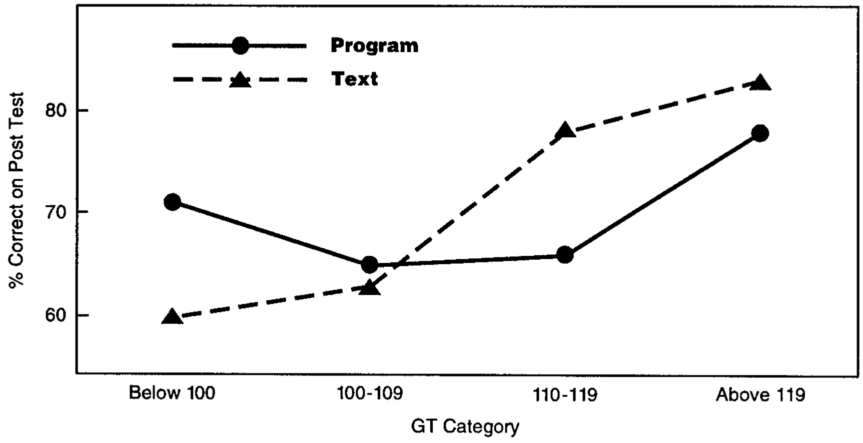
Note. R stands for random assignment to the Unaided Program and to the Text Version.

Table 5
Results With Infantry Trainees: Mean Percent Correct on Total Score

Experimental Conditions	Test	
	Posttest	Retention
Unaided Program (Demonstrations)	70	64
Text Version	68	60
Baseline	▨	46

Figure 3.

Posttest scores for the Program and Text groups for Infantry trainees with different GT scores [$F(3, 73) = 6.23, p < .0008$].



Interestingly, on both the posttest and the retention test, the two versions of the program had different effects on trainees with differing levels of ability, as assessed by the General Technical (GT) score from the Armed Services Vocational Aptitude Battery (ASVAB). TheGT score is a combination of verbal and arithmetic reasoning subtests from the ASVAB, and therefore, was assumed to provide a measure of general ability. The trainees with high GT scores benefited more from the text version than demonstration version of the program; trainees with low GT scores benefited more from the demonstration version than the text version (see Figure 3). These results suggested that the effectiveness of the different versions of the program was a function of the learners' strengths and weaknesses. Trainees with the higher GT scores profited from the text version which demanded their verbal and reading skills, whereas the demonstration version stressed auditory and perceptual skills. The reading skills of the trainees with the higher GT scores may have also been hampered in the demonstration version as some of the word slides were difficult to read, being set at 20/50 visual acuity to simulate reduced visual acuity at night. Apparently, the auditory and perceptual aspects of the demonstration version of the program compensated for the more limited reading skills of the trainees with the lower GT scores, whereas the text version did not.

Conclusions/Observations

The findings showed that soldiers' prior knowledge of unaided night vision tended to be fragmentary; they answered only half the test items correctly. Soldiers indicated little to no previous formal instruction on unaided night vision, which supported the front-end analysis findings of a lack of current training and instructional material in this area.

All evaluations showed the unaided program to be very effective and to reduce a training deficiency in the Army's current doctrine and training literature and training programs. Little forgetting occurred over a 3 week period. The text version was also effective, particularly for soldiers with high verbal ability. Success of the text version was attributed in part to the extensive work that had gone into developing and testing the script for the demonstration version of the program, as the text version was practically identical to the script.

In general, the success of the program was attributed to several factors: the use of ground force examples which facilitated understanding and increased interest, effective application of the 35-mm slide technology to the demonstrate critical concepts and perceptual phenomena at night, and the inadequacy of "on-the-job" training or field experience only in this domain.

Impact of Findings for Army Use

The findings with Infantry trainees suggested that both demonstration and text versions of the program may be needed to maximize learning for all soldiers. Consequently, a job aid highlighting basic concepts and guidelines, as well as a more detailed summary of program content, were developed (Dyer & Mittelman, 1995). These can serve as instructional guides prior to receiving the program, as a "memory jogger" after the program, or as both.

Substantial effort was put into making the training package easy to use. The program is configured as an exportable training package. All the necessary instructional materials and training aids are included. The package also includes an audio tape to help train instructors. This tape can substitute for an on-site instructor when a trained one is unavailable. However, the preferred instructional mode is with an experienced instructor, as this provides the best means of interacting with students and conducting the demonstrations.

Both experienced and inexperienced soldiers can profit from seeing the program. The knowledge gained can be applied directly by soldiers to maximize the use of their eyes at night and applied by leaders to refine their standing operating procedures for night operations. It is recommended that the program be repeated periodically to maximize retention, understanding, and application of unaided night vision principles and skills during the conduct of night operations.

Although the program is probably the most complete program on unaided night vision currently available, it does not train certain night vision skills such as the ability to use diamond viewing habitually under stress and fatigue, to estimate distance under different levels of illumination, or silhouette recognition. A different form of training would be required for such skills.

Implications for Future Research

The findings have implications for future research on the effectiveness of different training and instructional media, on how individuals understand and retain perceptual phenomena, and how to measure individuals' understanding of perceptual information. Such research is particularly important to the NIGHTFIGHTER research program, given that soldiers have distinct views of the battlefield at night when they use their unaided eye and night vision devices such as image intensification devices and thermal sights.

The experiment with Infantry trainees showed that no instructional medium is inherently better than another. The effectiveness of the medium varied with the aptitude of the trainee. The text version was better for trainees with high GT scores; the slide presentation was better for trainees with low GT scores. These results support the proposition that instructional techniques whose symbol systems correspond to the learners' strengths and which compensate for learner weaknesses make learning easier. Thus when investigating how to train soldiers to use night vision devices and to interpret the different images provided by these devices, there is a need to attend to possible individual differences in the ways soldiers acquire and retain perceptual information and to tailor the instructional media accordingly. New and/or sensitive measures of what is learned, how it is understood, and what is applied are also needed to obtain a complete picture of soldier proficiency.

Research Payoff to the Army

Commanders should expect a 30% to 40% increase in their soldiers' knowledge of night vision, knowledge and information which can be applied directly to improve soldier performance and to refine unit standing operating procedures for night operations. The program is being used by the opposing force at the Joint Readiness Training Center, the 82d Airborne Division, and the Ranger Training Brigade. The program will be part of the U.S. Army Infantry School's Dismounted Battlespace Battle Lab's Night Fighting Training Facility and available in the exportable training package to be distributed by the Battle Lab.

References

Dyer, J. L., Gaillard, K., McClure, N. R., & Osborn, S. M. (1995). *Evaluation of an unaided night vision instructional program for ground forces* (Technical Report). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. Manuscript in preparation.

Dyer, J. L., & Mittelman, M. H. (1995). *An unaided night vision instructional program for ground forces* (Research Product). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. Manuscript in preparation.

Mittelman, M. H., & Still, D. L. (1989). *Unaided night vision training guide*. Pensacola, FL: Naval Aerospace Medical Institute and Naval Aerospace Medical Research Laboratory.

Acknowledgements

The assistance of all the leaders, soldiers, instructors, and units who participated in evaluations of this program, whose recommendations were used to improve the program, and/or who supported program implementation must be acknowledged: the Ranger Training Brigade; the 3rd Battalion, 75th Ranger Regiment; the opposing force and observer/controllers at the Joint Readiness Training Center; the 2d Bde, 82d Airborne Division; and the Infantry Reception Station. The substantial contributions of Cdr Michael Mittelman of the Naval Aerospace Medical Research Laboratory must also be cited. His continuing support was essential to the production of the final ground forces' program.

How to Make Decisions About the Effectiveness of Device-Based Training: Elaborations on "What Everybody Knows"[1]

John A. Boldovici, Ph.D. and Eugenia M. Kolasinski, Ph.D.[2]
U.S. Army Research Institute
Simulator Systems Research Unit, Orlando

Abstract

Statisticians, biomedical researchers, and behavioral scientists have publicized errors in examinations of the differential effects of two or more treatments. The publicity about those errors seems to have been ignored by many applied behavioral researchers, including some responsible for evaluations of device-based training in the U.S. Army. Ignoring the causes and effects of the common evaluation errors, and especially errors associated with hypotheses of equal effectiveness of conventional training and device-based training, leads to logical contradictions, threats to readiness, and no scientifically legitimate ways to examine the effects of OPTEMPO alterations. Those problems may be avoided by applying a few basics of statistical analysis and inference to the design and interpretation of evaluations of device-based training. The basics comprise hypothesis tests, power analyses, and confidence intervals; they are elaborated in this paper with examples of how to apply each to designing and interpreting evaluations of the Army's forthcoming Close Combat Tactical Trainer.

[1] This paper summarizes some of the thoughts presented by the senior author, with assistance from the junior author, to the Special Advisor, Office of the Deputy Undersecretary of the Army (Operations Research) and to the U.S. Operational Evaluation Command as part of an agreement for advisory services between the U.S. Army Research Institute and COL James E. Shiflett, Project Manager for Combined Arms Tactical Training. We thank Stephen L. Goldberg, William C. Osborn, and Eric A. Vaden for their thoughtful and informed reviews. We are especially grateful to Peter W. Gieser and David W. Bessemer for their reviews and for their advice about hypothesis testing and statistical power throughout preparation of this paper.

[2] Address correspondence about this paper to either author at U.S. Army Research Institute, Simulator Systems Research Unit, 12350 Research Parkway, Orlando, FL 32862-3276. Send electronic mail to boldovij@stricom.army.mil or kolasing@stricom.army.mil.

Statisticians, biomedical researchers, and behavioral scientists have written about errors in experiments designed to examine the differential effects of two or more treatments. Blackwelder (1982), Boldovici (1987), Boldovici and Bessemer (1994), Burmeister (1992), Cohen (1962, 1990, 1994), Fisher (1942), Frick (1995) Gigerenzer (1993), and Lehmann (1959) have addressed the following errors and how the errors affect interpreting the outcomes of experiments: testing null hypotheses of equality of treatment effects, insufficient statistical power to detect differences between treatment effects, failure to specify beta risks and interpret null results in light of confidence intervals, and misinterpreting statistically non-significant differences to signify equal effectiveness of treatments. Widespread knowledge about those errors[3] seems to have had, as the data in Table 1 suggest, no noticeable effect on the course of recent human-factors and behavioral-science research.

Table 1

Numbers of Power-Analysis (PA) and Confidence-Interval (CI) Reports in Three Professional Journals

	Human Factors	*J Appl Psychol*	*JAMA*
1994 Months	Jan-Dec	Jan-Jun[a]	Jan-Aug[a]
Tot Articles	49	41	109[b]
Articles NA[c]	5	4	0
Nbr PA Repts	0	0	18
Nbr CI Repts	0	0	66

[a]Inclusion of additional 1994 issues, which were not available from our source, would not have affected the point made by these data. [b]Two special issues were excluded: *Contempo* and *Peer Review*, which contained no reports of experiments. [c]Articles not applicable were non-experimental; e.g., theories, literature reviews, statistics, including one article in *J Appl Psychol* that compared the statistical power of two procedures.

[3]Cohen (1994), Rozeboom (1960), and others noted there was nothing new about the errors addressed in their articles. Bakan (1966) wrote that discussing "mischief . . . associated with the test of significance [is]. . . hardly original [and is] 'what everybody knows'. . .[but to] say it out loud is to assume the role of the child who pointed out that the emperor was outfitted in his underwear" (p. 423). The present authors admit too that there is nothing new about the errors and solutions we are about to discuss. We hope, however, that our suggestions about potential effects of compromised evaluations on readiness and downsizing will have value for evaluation designers and policy makers.

The reasons for behavioral researchers' ignoring the admonitions of the authors cited above are hard to discern. A review of recent issues of the *Journal of the American Medical Association* (see Table 1) suggested, however, that those admonitions are more likely to be heeded by biomedical researchers than by behavioral researchers, perhaps because the life-or-death implications of applied biomedical work are more obvious than are the life-or-death implications of applied behavioral research.

In the realm of applied behavioral research, the admonitions of the authors mentioned above seem often ignored in evaluations of Army training devices. Our experience in advising the Army in planning evaluations of device-based training is that compromised evaluations usually are rationalized along the following lines: "We may not have sufficient statistical power to detect statistically significant differences between the scores of compared groups, but the test results will 'at least put us in the ball park.'" Nothing could be farther from the truth: We have no a priori criteria for judging whether we are in the ball park; this is an issue of generality of results whose resolution requires replication, which is not feasible for multi-million-dollar tests of device-based training. The ball park, like many so-called 80% solutions, is defined after the fact as wherever the results happen to put us. Believing that low-power tests of device-based training will "at least put us in the ball park" is unfounded and flies in the face of basic statistics: If we conduct compromised device tests and find no statistically significant differences between the scores of compared groups (e.g., conventionally trained vs. device-trained), then the results are, contrary to the ball-park thinking, no better than guessing: Random or error variance exceeded that of treatment effects, and the test might as well not have been conducted. That is especially true for cases in which we suspected or knew in advance that the power of our device test was so weak as to preclude finding statistically significant differences between compared groups' scores.

The ball-park line of thinking disconcerts additionally because null results in military device evaluations may be taken, without supporting analyses, as evidence that conventional training and device-based training are equally effective. Null results in training evaluations can of course ensue from causes other than equal effectiveness of the compared training. Those causes were reported by Orlansky, Dahlman, Hammon, Metzko, Taylor, and Youngblut (1994) in the context of Simulation Networking (SIMNET) evaluations and include small sample sizes, inadequate test designs, and other evaluation deficiencies.

Rationale

Several reasons underlie our concern about misinterpreting null results to signify equal effectiveness of conventional training and device-based training. On a logical level we find the notion untenable that field training and device-based training are equally effective—as Army leadership apparently does too. The Army's concern with developing effective mixes of field training and device training belies the equivalence of field training and device-based training. If field training and device training were equally effective, then decisions about training strategies would be based on price alone; the medium wouldn't matter.

The illogic of equal effectiveness also is apparent from reading about or watching field training and device training: Field training is more effective than device training for some tasks, and device training is more effective than field training for other tasks. The two kinds of training cannot therefore be equally effective and can only be shown to be equally effective in one or both of two ways: (1) by using evaluation designs, performance measures, and analysis methods so insensitive as to fail to detect differences visible to the naked eye, and (2) by misinterpreting null results.

More important than our short-term concerns about logical contradictions are the longer term implications of the equal-effectiveness myth for downsizing and readiness. As Boldovici and Bessemer (1994) showed, evaluation designs that yield findings of no difference between the effects of field training and device training almost always contain fatal flaws, that is, flaws so severe as to preclude finding differences that in fact exist. If one were to use similarly flawed evaluation designs to compare, for example, sustainment training and no sustainment training, the evaluations would yield null results for the same reasons—insufficient statistical power and other design flaws—that comparisons of field training and device training yield null results. Downsizers may as legitimately use null results to tout equal effectiveness of training and no training as device advocates use null results to tout equal effectiveness of field training and device training.

In addition to providing precedent for spuriously demonstrating the equivalence of training and no training, the myth of equal effectiveness of field training and device training paves the way for closing training and maneuver areas and for additional decreases in resources that attend field training. Downsizers' contentions are easy to foresee: "If device training and field training are equally effective, then what harm can come from additional substitutions of device training for field training, that is, from additional reductions in OPTEMPO?"[4] The flaws in that line of thinking can be exposed by applying legitimate methods for examining the equivalence (and non-equivalence) of alternative kinds of training—methods which we shall discuss shortly and which, to the best of our knowledge, have not been used in evaluations of device-based training in the Army. Military leaders and the device evaluators who advise military leaders need to understand the differences between legitimate and illegitimate methods for establishing the equivalence of alternative kinds and amounts of training. That understanding is essential to ensuring the use of legitimate methods for examining the effects of device-based training and of OPTEMPO alterations.

Our final reason for concern with misinterpreting null results to signify equal effectiveness of field training and device training is as Jack H. Hiller (personal communication, August 1994) noted: How will readiness be affected by military doctrine and training that are based on assumptions about equal effectiveness if those assumptions are wrong? If training with devices is less effective than field training, as it surely is in many cases, then claims of equal effectiveness provide untenable bases for sustaining readiness. Hiller's thinking suggests that device evaluators should be as concerned about errors in examining the equivalence of training regimens as biomedical researchers are about errors in examining the equivalence of pharmacologic treatments: In both cases evaluation results factor into life-or-death decisions.

Overview

Avoiding the logical errors and threats to readiness summarized above requires understanding and applying a few basics of statistical analysis and inference, which Bakan (1966) called "what everybody knows." We shall elaborate three aspects of what everybody knows—hypothesis tests, statistical power, and confidence intervals—in the context of evaluating the Army's forthcoming Close Combat Tactical Trainer (CCTT).

[4] OPTEMPO is the Army's abbreviation for operating tempo; it refers to "the annual operating miles or hours for the major equipment system in a battalion-level or equivalent organization" (National Simulation Center, 1994).

Hypothesis Tests

Tests of Army training devices typically compare the effects of conventional or field training to an altered training regimen in which part of the conventional or field training is replaced by device-based training.[5] Because device-based training may be proposed to replace some parts of conventional training, a question naturally arises about whether the proposed substitution will adversely affect soldiers' proficiency as compared to the proficiency of soldiers who train with existing, conventional means. That question easily translates to a null hypothesis of equality of treatment effects, H_O: $\mu_c = \mu_d$ (where μ_c and μ_d are the mean scores of the conventional and device groups) and may be formulated as such by device evaluators. The problem with stating comparisons in terms of no difference between treatment effects is as R. A. Fisher noted in 1942:

> The null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis (p. 16).

The easy way to avoid the problem implied by Fisher, that is, erroneous acceptance of H_O (Type II error), is as we were taught by our early statistics instructors: Never accept H_O, and thus avoid the possibility of accepting H_O erroneously. That lesson, although logically irrefutable, is intellectually unsatisfying because null results immediately engender (we hope) the question, "Did we find no differences because there are no differences or because flaws in our experiment precluded finding differences?"

A more satisfying way to assess null results than to dismiss all out of hand is to specify β, which is the probability of Type II error. Specification of β is contingent upon a specific treatment difference (i.e., β covers only one specific treatment difference). By specifying a value for β, evaluators, researchers, or policy makers set the risk they are willing to take of making a Type II error for a specified effect. The utility of specifying β can be seen by imagining how our interpre-

[5] For convenience we shall refer to these alternatives as conventional training and device-based training.

tation of a null result would differ with, say, $\beta = .80$ or $\beta = .20$: With $\beta = .20$ our temptation to base decisions on the evaluation result would be greater than with $\beta = .80$.

Morrison (1990), citing Kirk (1984), noted that researchers

> . . . have suggested that a ratio of 4 to 1 (Type II to Type I error probabilities) is an acceptable relationship between the two types of error. Using [the widely accepted] .05 value for α implies that .20 is an acceptable value for β. Because β is the complement of power (1 -β), the commonly accepted value for power is then .80 (p. 12).

Power = .80 and $\beta = .20$ are not, of course, mandatory values any more than is $\alpha = .05$. Policy makers can adjust those values depending on the importance (as defined by costs and hazards) of errors in decisions that will ensue from the evaluation results.

Specifying β allows us to accept H_O the same way we routinely accept H_A, that is, with the understanding that in approximately $\beta(100)$ times in 100, for a given effect size, we will be wrong. With β and α specified, evaluators can avoid the error of automatically equating null results with equal effectiveness by stating their conclusion in the following general form: "With $\alpha = X$, $\beta = Y$, and effect size = Z, we found no statistically significant differences between the compared groups' scores." Presenting null results in other than that general form is in our view a disservice to the evaluation customer and invites suspicion of playing fast and loose with the data.

There are at least two additional ways to address Fisher's concern about accepting null hypotheses. One is by using confidence intervals and will be discussed later. The other is as discussed by Blackwelder (1982) and by Rosenthal and Rubin (1994), who recommended specifying H_O and H_A so that, "Type I error α and Type II error β are reversed from the case of the usual null hypothesis" (Blackwelder, 1982, p. 349). This rearrangement leads to testing the null hypothesis that the standard treatment (conventional training in our case) is more effective than the experimental treatment (device-based training) by a specified amount, δ. Rejecting H_O, that conventional training is more effective than device-based training by δ or more, and accepting H_A, that conventional training and device training differ by less than δ, are conclusions with which behavioral researchers are likely to be comfortable and which, because α is routinely specified, are consistent with traditional hypothesis testing (Blackwelder, 1982).

Statistical Power

The power of a statistical test is the probability that the test will find an effect (a difference between the mean scores of compared groups in our case) given that an effect of a certain size exists. Without sufficient power, real differences between the proficiency of conventionally trained and device-trained groups will go undetected. Power is a function of three quantities: (1) sample size, (2) variance between and within compared groups' scores, and (3) effect size, that is, the size of the actual difference between compared groups' scores.

Selecting sample sizes that are neither so small as to preclude finding differences between compared groups' scores nor so large as to waste evaluation resources is a straightforward matter whose implementation can save money: If on the one hand, multi-million dollar device tests are planned and the power of the tests is unknown (as is the case for the CCTT), then we may waste the entire cost of the test. If the power of the test is computed and found to be too weak to reveal existing differences between the scores of the compared groups, and the test is conducted as-is, then we certainly waste the entire cost of the test. If on the other hand, the power of a device test is computed and found to be in the mid- to high-nineties with a sample size in the zone of diminishing returns on power, then policy makers may choose to reduce the sample size and save the attendant costs.

The costs of scrimping or squandering sample sizes increase as the focus of device testing moves up echelons, from individual crewmen or tank-commander-gunner pairs in tank-gunnery trainers, through crews and platoons for SIMNET, to companies and eventually battalions for the CCTT. That is because sample-size requirements remain the same regardless of which echelon is used as the sampling unit. Conducting individual- or crew-level tests with insufficient power to detect differences between compared groups' scores may be viewed as a negligible waste of evaluation resources. Conducting similar tests to compare groups of companies or battalions may be viewed as unconscionable.

Costly errors such as those hypothesized above may be avoided by doing power analyses before comparing the effects of conventional and device-based training. Results of the power analyses will tell us, with given sample sizes, the probability of finding differences that exist between the scores of compared groups. With knowledge of the probability of finding true differences between the compared groups' scores, we can make informed decisions about whether to spend the money to conduct the tests. Consider, for example, how our decision about whether to conduct a comparison between device-based training and conventional training might differ depending on whether the power analyses told us we had a 5% chance or a 95% chance of detecting real differences between compared groups' scores. Such informed decisions have, to the best of our knowledge, never been made in planning evaluations for Army training devices; the power analyses were not done.[6]

Examples of Power Computations

Assume for purposes of illustration that an evaluation is planned for the CCTT: Difference scores between pre- and post-tests will be compared for two battalion-size task forces, one of which trains conventionally and the other of which replaces some part of conventional training with CCTT training. Assume further that a two-sample *t* test will be used to examine the differences between the two groups' mean scores and that, because the CCTT is a company-team training device, analyses will be conducted with companies as the sampling units. Thus each of the compared task forces will comprise four companies. H_O is that $\mu_c = \mu_d$, and H_A is $\mu_c > \mu_d$. In this case the test statistic is,

$$T = \frac{\overline{x}_c - \overline{x}_d}{\sqrt{\dfrac{s_c^2 + s_d^2}{n}}}$$

where *n* is the number of observations per group, assumed to be equal and in our case 4. H_O will be rejected if $T > t_{(\alpha,2(n-1))}$, where $t_{(\alpha,2(n-1))}$ is the upper-tail α percentage point of the Studentized *t* distribution with 2(n-1) degrees of freedom. The power of this test is given by,

$$1 - \phi(t_{(\alpha,2(n-1))} \mid 2(n-1), \frac{\Delta}{\sigma} * \sqrt{n/2})$$

which is the upper-tail probability for the $t_{(\alpha,2(n-1))}$ percentage point of the Noncentral *t* distribution with noncentrality parameter ($\Delta/\sigma*(n/2)^{1/2}$) and 2(n-1) degrees of freedom.

[6] As is the case with proving H_O, we realize the impossibility of proving that no power analyses were done. We welcome evidence to the contrary.

Computing the statistical power of the CCTT test hypothesized above requires (1) a sample-size estimate (given earlier as four companies per task force), (2) an effect-size specification by the evaluation proponent or customer, and (3) a variance estimate, most conveniently but not necessarily, obtained from a related evaluation.

One method of computing the power of the CCTT test hypothesized above involves using data from a related evaluation to estimate variance. First let's assume that effect sizes of 10% and 20% are of interest to the evaluation proponent; that is, the evaluation proponent is willing to live with a 10% difference in favor of conventional training over CCTT training, but the proponent believes a 20% difference in favor of conventional training is wholly unacceptable: The 20% difference will, for example, require devising and implementing entirely new training strategies. Using data from a SIMNET evaluation by Brown, Pishel, and Southard (1988)[7] in the form of a two-sample *t* test shows effect sizes of 10% and 20% to be 2.1 and 4.2, and an estimate of the population variance to be 10.5 as given by,

$$\mathrm{s}_\mathrm{p}^2 = \frac{(n_\mathrm{c} - 1)\,\mathrm{s}_\mathrm{c}^2 + (n_\mathrm{d} - 1)\,\mathrm{s}_\mathrm{d}^2}{n_\mathrm{c} + n_\mathrm{d} - 2}$$

Using the estimates of effect size = 2.1, variance = 10.5, and $\alpha = .05$, we find power ≈ .20 with $n = 4$ companies per group. Using those same estimates of variance, α, and n, but with the 20% effect size = 4.2, we find power ≈ .49. That is, with $n = 4$ companies per group, we run an 80% risk of failing to detect a 10% difference between compared groups' scores and a 51% risk of failing to detect a 20% difference (given as unacceptable in our example).

Should the proponents of our hypothetical CCTT evaluation choose to spend several million additional dollars they could double the number

[7] Brown, Pishel, and Southard (1988) compared the scores of two 4-platoon experimental groups, one of which trained conventionally and the other of which used SIMNET in training. Both groups performed various tasks before and after training. Ten of the tasks were performed in common by all eight platoons. Each platoon was scored GO or NO-GO on the ten tasks before and after training. We computed differences between Brown et al.'s pre- and post-training scores and used the difference scores as the dependent measure, thus making the design a two-sample *t* test. We then used the difference scores to estimate variance in our examples of the power analyses and confidence intervals presented in this paper.

of companies in the evaluation to eight per compared group. Doing so would yield power ≈ .34 for detecting the 10% difference, and power ≈ .80 for detecting the 20% difference. And if our hypothetical proponents decided to spend several more millions of dollars they could triple the number of companies to 12 per group. Doing so would yield power ≈ .46 for the 10% difference and power ≈ .92 for the 20% difference. Salient features of this example include: (1) Even by tripling *n* to 12 companies per group, we still have greater than a 50% chance of failing to detect real differences of as much as 10% between the compared groups' scores. (2) To detect differences as large as 20% we shall need either 8 or 12 companies per group depending upon whether we are satisfied with a 20% chance or an 8% chance of failing to find differences that do in fact exist. (3) In no case do four companies per group suffice.

Another way to compute statistical power requires no estimates or computations of variance, but only an assumption about the compared groups' scores in terms of standard-deviation units. Suppose, for example, we decide that a difference between compared groups' scores of one standard deviation is meaningful. With the underlying effect-size:standard-deviation ratio (Δ/σ in the power formula) of 1.0, and using $n = 4$ companies per task force, we find power ≈ .35. Doubling and tripling *n* yield powers ≈ .60 and .77. In this case neither four nor eight companies per group can satisfy the power requirement. And with 12 companies per compared group we still have a 23% chance of failing to find true differences.

The methods of power analysis summarized above take numbers of observations as given and compute power based on the given numbers of observations. Another way to use power analyses is to prescribe β as discussed earlier and to let the power analysis tell us what sample size will satisfy our prescribed β. If, for example, we're willing to take only a 5% chance of accepting a null result erroneously, then the evaluation is likely to require far greater *n* than if we're willing to take a 50% or a 95% chance of the erroneous decision.[8] For the CCTT evaluation example hypothesized above, Table 2 gives information for determining the

[8] With $\beta = .95$, the power of the test is only .05. Gigerenzer (1993) reported that a similar situation led Neyman (1950) to note that some of R. A. Fisher's tests were "worse than useless," because their power was less than their size (α).

minimum sample sizes necessary to satisfy β = .05, β = .50, and the absurd extreme of β = .95, which correspond to powers of .95, .50, and .05, respectively. Effect sizes are 10% or 20%, and the variance estimate is computed from Brown et al.'s (1988) data. The data in Table 2 show that with effect size = 10% and variance = 10.5, 52 companies are required to satisfy β = .05 and 14 companies for β ≈ .50. The absurd extreme of β = .95 is not possible because 2, the minimum number of companies necessary for statistical inference, yields β = .88. With effect size = 20%, the closest approximation for β = .05 can be obtained with 14 companies, and the closest approximation for β = .50 can be obtained with 5 companies. The extreme β = .95 is again not possible, because using only 2 companies yields β = .77.

The utility of computing *n* for various betas is in its fairness to customers, that is, taxpayers at large and the military leaders who decide how to spend device-evaluation money. Those customers would, we suspect, make different decisions about conducting device evaluations depending on whether finding no statistically significant difference between the scores of compared groups carried a 5% or a 95% risk of being wrong. Needless to say, making such decisions requires knowing the numbers.

Table 2
Minimum Numbers of Companies (n) *Necessary to Satisfy ß = .05, .50, and .95 With Effect Sizes = 10% and 20% and Variance = 10.5*

Effect Size	Power	ß	n[a]
2.1 (= 10% x 21)	.95	.05	52
	.51	.49	14
	.12	.88	2
4.2 (= 20% x 21)	.96	.04	14
	.59	.41	5
	.23	.77	2

[a] Sample sizes are the lowest values of *n* that exceed the desired power level.

Confidence Intervals

Confidence intervals differ from hypothesis tests but are closely related.[9] The confidence interval is formed from a combination of the statistic, the variance of the statistic, and the critical value to which the test statistic is compared. If the same α is used, then the decision to reject or not to reject H_O will be the same whether a confidence interval or a hypothesis test is used. The advantage of using confidence intervals is that, in addition to permitting hypothesis testing, confidence intervals provide bounds on an evaluation's effect-size estimate, that is, bounds on the observed difference between the compared groups' mean scores: "A hypothesis test tells us whether the observed data are consistent with the null hypothesis, and a confidence interval tells us which hypotheses are consistent with the data" (Blackwelder, 1982, p. 350). That is, the confidence interval displays the set of differences that are plausible given the data obtained. For a 100(1-α)% confidence interval, the conclusion is, "We can be 100(1-α)% confident that the interval contains the true value of the difference between the compared groups' mean scores." A wide confidence interval, as compared to a narrow confidence interval, indicates that a greater proportion of the range of possible differences between the compared groups' mean scores is included in the interval: The narrow confidence interval indicates fewer possible values for the difference between compared groups' means than does a wide confidence interval.

Examples of Confidence Intervals

In our reanalysis of Brown et al.'s (1988) data as a two-sample *t* test, the means and standard deviations for the SIMNET and conventional groups were, respectively, 3.0 (s = 2.94) and -.25 (s = 3.5). These means do not differ statistically, as shown with a two-tailed, two-sample *t* test with $\alpha = .05$ ($t = 1.422$, p = .204). A two-sided confidence interval for the difference in means is (-2.343, +8.843), as given by,

$$(\bar{x}_c - \bar{x}_d) \pm t_{\frac{\alpha}{2}} \sqrt{s_p^2 \left(\frac{1}{n_c} + \frac{1}{n_d}\right)}$$

[9] Blackwelder (1982) pointed out that, "Although the theory of hypothesis testing is useful particularly in planning a clinical trial [device evaluation in our case], the confidence interval approach may be more useful in the analysis, interpretation, and reporting of the accumulated data . . ." (p. 350).

where s_p^2 is as defined earlier and $t_{\alpha/2}$ is based on ($n_c + n_d$ - 2) degrees of freedom.

Confidence intervals have two important characteristics. The first important characteristic is, as implied above, whether the confidence interval contains zero. If the interval does not contain zero, then the null hypothesis of equality, H_O: $\mu_c = \mu_d$[10], must be rejected; that is, the difference between the compared groups' mean scores is statistically significant. The confidence interval for Brown et al.'s data (-2.343, +8.843) contains zero. So with $\alpha = .05$ we decide not to reject H_O: $\mu_c = \mu_d$, which is the same decision we reached with the two-tailed hypothesis test and the same α.

The second important characteristic of confidence intervals is their width. A narrow confidence interval gives more precise information about the location of the difference between compared groups' mean scores than does a wide interval. If the interval for Brown et al.'s data were, for example (-1, +1), we could make a better guess about where the difference between means lies than we can make given the actual confidence interval (-2.343, +8.843): Equal effectiveness of the compared training regimens would be more plausible with the narrow interval than with the wide interval. The confidence interval for Brown et al.'s data is wide: At greater than 11.0 Brown et al.'s interval includes over half the range of possible differences between the compared groups' means, which is -10.0 to +10.0. Brown et al.'s interval does contain zero though, so we must not dismiss an equal-effectiveness interpretation of the evaluation result. But in addition to zero the interval contains every other possible difference from -2.343 to +8.843. Concluding equal effectiveness correctly is therefore less likely than it would be if the interval were narrower. Thus both the hypothesis test and the confidence interval led us not to reject the possibility of equal effectiveness. But the confidence interval provided additional information suggesting a high degree of uncertainty associated with an equal-effectiveness interpretation.

[10] The example we are using is for null hypotheses of equality, that is, H_O:$\mu_c = \mu_d$. The logic of confidence intervals also applies to null hypotheses other than of equality, that is, H_O:$\mu_c \geq \mu_d$, which as noted earlier may be desirable.

Summary

Methods were presented for avoiding the logical contradictions and threats to readiness that attend misinterpreting evaluation results to signify equal effectiveness of conventional training and device-based training. Our methods are elaborations of what Bakan (1966) called "what everybody knows." Applying these elaborations of what everybody knows is essential to scientifically legitimate examinations of device-based training and of OPTEMPO alternatives. Our elaborations of what everybody knows reduce to six prescripts:

1. Specify a value for β indicating the risk the evaluation proponent is willing to take of erroneously accepting the null hypothesis of equality, that is, of making a Type II error.

2. Perform analyses to determine the power of tests with n given or to determine what n must be to satisfy various values of β, including the maximum value of β the evaluation proponent is willing to accept. Supplying effect-size estimates for performing power analyses is the evaluation proponent's responsibility. Variance estimates for power analyses can be obtained from data in related evaluations. If variance estimates, effect-size estimates, or both are not available, then use an effect-size:standard-deviation ratio (Δ/σ).

3. Specify H_O and H_A such that the roles of Type I error and Type II error are reversed. Using hypotheses in the form H_O: $\mu_c \geq \mu_d + \delta$ and H_A: $\mu_c < \mu_d + \delta$ leads to conclusions with which behavioral researchers are likely to be comfortable and which, because α is routinely specified, are consistent with traditional hypothesis testing.

4. Report evaluation results with α and β specified.

5. Avoid the possibility of erroneously accepting a null hypothesis of equality by never accepting a null hypothesis of equality.

6. Use confidence intervals to set bounds on the between-group difference estimate and to help recognize the extent of possible error in an interpretation of equal effectiveness.

References

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66*, 423-437.

Blackwelder, W. C. (1982). "Proving the null hypothesis" in clinical trials. *Controlled Clinical Trials, 3*, 345-353.

Boldovici, J. A. (1987). Measuring transfer in military settings. In S. M. Cormier & J. D. Hagman (Eds.), *Transfer of learning* (pp. 239-260). Orlando, FL: Academic Press.

Boldovici, J. A., & Bessemer, D. W. (1994). *Training research with distributed interactive simulations: Lessons learned from simulation networking.* (ARI Technical Report 1006). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A285 584)

Brown, R. E., Pishel, R. E., & Southard, L. D. (1988, April). *Simulation Networking (SIMNET) preliminary training developments study (PTDS)* (TRAC-WSMR-TEA-8-88). White Sands Missile Range, NM: U.S. Army TRADOC Analysis Command.

Burmeister, L. F. (1992). Proving the null hypothesis. *Infection Control and Hospital Epidemiology*, 13(1), 55-57.

Cohen, J. (1994). The earth is round (p<.05). *American Psychologist*, 49(12), 997-1003.

Cohen J. (1990). Things I have learned (so far). *American Psychologist*, 45(12), 1304-1312.

Cohen, J. (1962). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Fisher, R. A. (1942). *The design of experiments.* London: Oliver and Boyd.

Frick, R. W. (1995). Accepting the null hypothesis. *Memory and Cognition, 23*(1), 132-138.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: methodological issues* (pp. 311-339). Hillsdale, NJ: Erlbaum.

Kirk, R. E. (1984). *Elementary statistics* (2nd ed.). Monterey, CA: Brooks/Cole Publishing.

Lehmann, E. L. (1959). *Testing statistical hypotheses.* New York: Wiley.

Morrison, J. E. (1990). *Power analysis of gunnery performance measures: Differences between means of two independent groups* (ARI Technical Report 872). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A219 917)

National Simulation Center. (1994). *Training with simulations: A handbook for commanders and trainers.* Fort Leavenworth, KS: Author.

Neyman, J. (1950). *First course in probability and statistics.* New York: Holt.

Orlansky, J., Dahlman, C. J., Hammon, C. P., Metzko, J., Taylor, H. L., & Youngblut, C. (1994). *The value of simulation for training* (IDA Paper P-2982). Alexandria, VA: Institute for Defense Analysis.

Rosenthal, R., & Rubin, D. (1994). The counternull value of an effect size: A new statistic. *Psychological Science*, 5(6), 329-334.

Rozeboom. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin, 57*, 416-428.

Understanding and Improving Tactical Problem Solving

Jon J. Fallesen, Ph.D.
U.S. Army Research Institute
Ft. Leavenworth Research Unit

Abstract

Skillful problem solving is highly valued for tactical planning and conduct of battle. Yet the development of competency is somewhat of an enigma. Despite great interest, there is much uncertainty about what leads to good problem solving. The recent trend to examine naturalistic behavior is promising, but has yet to identify how leaders' styles differ and how effectiveness may vary. The current research identifies basic strategies and explores ways of teaching cognitive skills for problem solving.

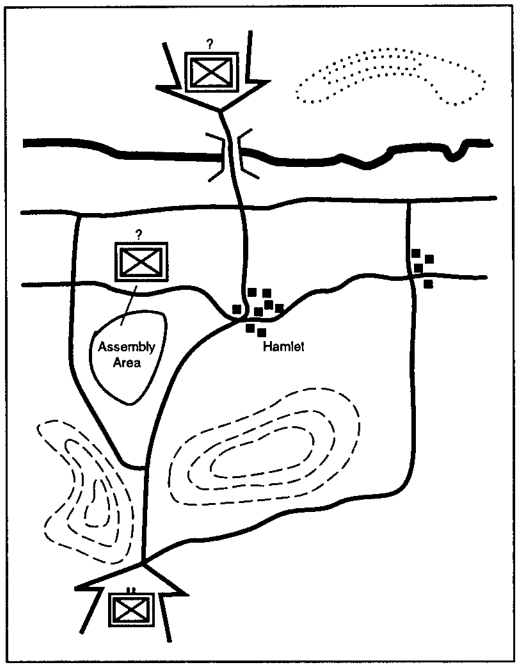
Battle commanders and their staffs do not always follow the tactical decision making model as taught. The lack of good measures to describe what problem solvers actually do makes it difficult to gauge the extent of the discrepancy. A catalog of problem solving strategies was developed as a basis of measurement. The baseline set was examined with a sample of staff officers. They rated the use and importance of 29 information processing and 19 choice strategies on three tactical problems. Results corresponded with expectations that indicate strategies are tailored to experience and situational constraints; however, strategies did not fall along either pure traditional or naturalistic lines. Emerging results have been used to develop cognitive skills instruction for a battle command course in the Command and General Staff College.

The Problem of Problem Solving

A good way to begin thinking about how our Army's leaders might solve problems is to consider a tactical situation.

Imagine that you are the Commander of 2-6 Infantry, 1st Brigade. The last you knew, friendly forces held a key bridge and the river line. Tomorrow the Division begins a major offensive across the river, with your battalion spearheading the attack starting at 0400. Right now it is 2000 and you are moving north to an assembly area as shown in the diagram on the next page. On the way you receive reports that *enemy infantry is occupying your assembly area and is continuing to reinforce. The size is unknown but estimated to be at least a company. Another surprising report indicates there is no sign of friendly forces at the river or the bridge. What do you do?*

The immediate reaction may be to consider ways to clear the enemy from the assembly area. The Commander might consider establishing overwatch positions on the high ground south of the assembly area and then moving up the road with mechanized units. Or the Commander might want to swing around and attack from the west. Some Commanders might stop their road march and await instructions from Brigade before any further actions or plans. Other problem solvers might think that the situation at the bridge is more important than at the assembly area. They might want to secure the bridge right away, however, running the risk of tipping off the Division's plans too early. There are many ways of defining what the problem is. This tactical situation is used in the current research to gain a better understanding of actual problem solving strategies.



Doctrinal and instructional guidance suggests that an analytical approach to problem solving is best: that a problem solver should generate multiple courses of action, assess each independently, and then compare them. The Commander faced with the above situation could follow this "best" approach, but could apply it to a narrowly defined problem, such as ridding the enemy from the assembly area. Instead of following these careful steps, many problem solvers follow the natural process of their thoughts. This naturalistic, experiential based approach is not easily described by general-purpose steps. Rather than emphasizing steps, a naturalistic approach focuses on the application of knowledge to understand the situation and to determine what to do in that situation. The field of problem solving has been so focused on finding ways to teach people to use normative procedures or on systems and procedures to help people follow normative procedures that the implications of studying actual problem solving behaviors are still unclear.

If it is unclear what competent leaders do to solve tactical problems and how less-experienced leaders differ, then doctrine and instruction should be cautious about the guidance that is offered to solve problems. Guidance about how to make a choice can provide a false sense of confidence or sufficiency, if the critical problem is to understand what problem to solve and not what is the optimal option. This is just one example of many possible disconnects between teaching points based on the classical model and the actual complexities of solving problems. Explicit information about actual tendencies of tactical problem solvers is scarce. We need to understand the intricacies of problem solving more fully and develop and guide leaders accordingly. As an instructional and research community we have relied on the application of the classical model too long as the basis for improving decision making. If actual performance can be better understood, then better guidance can be developed for solving problems (Essens, Fallesen, McCann, Cannon-Bowers, & Dorfel, 1995).

Comparison of Classical and Naturalistic Problem Solving

Consider two classes of models that address problem solving or decision making. One set, referred to here as classical, focuses on the comparison of options. The classical approach attempts to produce the optimal outcome. Option comparisons are to be done in a concurrent fashion and are to be comprehensive in both number of options considered and the attributes on which options are assessed and compared. The control of the procedure comes from the external guidance provided in the model.

In contrast to this classical approach is one described as naturalistic. The focus of the naturalistic approach is on achieving an adequate situation understanding so the right problem is considered. Multiple options are considered only as necessary, so options occur sequentially rather than at the same time. The classical model provides little guidance about the source of options. The naturalistic model portrays options as a recognition of what is familiar or a process of finding what will work. The naturalistic procedures are not controlled in the same

Table 1
Comparison of Classical and Naturalistic Models Procedures

Features	Classical	Naturalistic
Control	Model Prescribed	Experience or determine during solution
Option source	Given or minimal guidance	Familiar or exploration
Option comparison	Concurrent	Sequential, if needed
Comparison factors	Comprehensive	Selective
Approach focus	Compare options	Understand situation, elaborate, improve
Outcome intention	"Optimal"	"Satisficing"

sense as classical ones. Naturalistic procedures are self-determined based on experience and depend on knowledge that is determined during problem solving. The naturalistic approach recognizes that it is difficult to prescribe in advance what procedure will efficiently and effectively derive a solution. After all, it is thinking that leads to the new knowledge necessary for resolution of the problem.

The implications for improving decision making from the two models are different. The classical model leads to teaching a systematic procedure that will supply the problem solver with the capability to

Table 2
Comparison of Classical and Naturalistic Models Training & Efficacy

Features	Classical	Naturalistic
Training	Systematic procedure Quantitative, analytical	Experience & feedback Thinking & monitoring
Model—recognized shortcoming	Don't use method	Lack experience
Others' criticism	Extensive effort Insufficient guidance False precision Not useful	Decisions not optimal Unreflective No simple description Too radical

meet future problems. It asserts that following an explicit procedure will lead to the best outcome. The naturalistic model leads to gaining experience to prepare for future problems. Further, if specific experience does not directly apply, then thinking and reasoning based on the similarities and differences from previous experience (and knowledge) need to provide the necessary adaptation.

Clearly military instruction and leader development deal with both procedural and experience-based instruction. The real issue may be that the two are not integrated to the extent that they should be, and when it comes to explicit guidance it is easier to hold up the "straightforward" steps of the classical model as the desired approach. Problem solving procedures are currently taught in officer basic, advanced, CAS3, and CGSOC instruction. An important question is whether the procedures that are taught are sufficiently adaptive to the richness of real situations.

Review of Tactical Problem Solving Findings

A review was conducted to gauge the existence of problems in tactical planning, attempting to understand the use of classical problem solving guidance (Fallesen, 1993). The general observation was that the classical model, embodied in the command estimate procedures, is not followed closely. The review sources consisted of studies of actual combat problem solving, records of NTC, JRTC, and BCTP rotations, surveys of commanders, and 30 military decision making experiments.

Table 3
Tactical Planning Performance Estimate Procedures

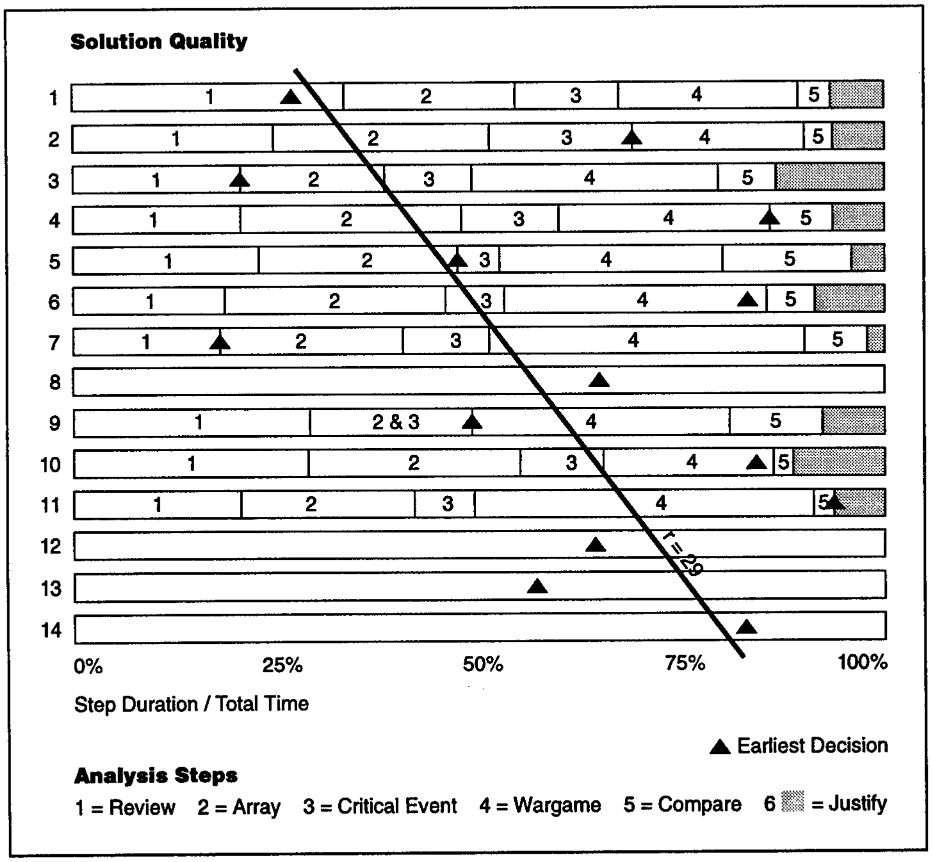
Failure to follow procedures
Single COA developed, single individual performs, leave out steps, poor navigation

Procedures are imprecise
Too formal, too much time, procedures are not cohesive

Excessive time demand
Not clear how to abbreviate

Inflexibility
Not clear how to change midstream, procedures lead to rigid, standard response

Discrepancies from the classical teachings and performance weaknesses occurred in many forms. There were cases where only single courses of action were developed, or alternate courses of action were developed that were not of genuine interest (sometimes called the "look-alike" and the "throw-away"). Steps are commonly left out, procedures do not help resolve many real issues (e.g., what to do when a mission change is inferred during a dynamic situation), or procedures do not identify a single best option. For instance, some studies showed that specific guidance (e.g., do not compare options until each has been independently assessed) provides little or no help or even interferes with determining good solutions (Fallesen, 1995). The above figure shows there is no significant correlation (r=.29, p=.16) between the earliest point a decision is made (depicted by triangles) and the quality of solution. The figure also shows that many decisions occurred much earlier than the comparison step that is specified in doctrine.



This review along with conceptual work on using the naturalistic approach for specifying procedures (Fallesen, Lussier, & Michel, 1992) contributed to the modification of doctrine for the command estimate that now specifies three different processes depending on the time available and the experience of the staff. However, the primary basis for the doctrinal process is still the classical model.

The review indicated that leaders do not closely follow the procedural guidance that is offered. So what do they do instead? Is what they do effective? If so, how can that be used to improve problem solving? If what they do is not effective, what else can be done to improve problem solving? One difficulty that the review highlighted was considerable variability in the way command and control was performed. Because of the variability it was difficult to measure styles and make comparisons. This conclusion led to the development of the following program of research.

Problem Solving Strategies

The current research was conceived as a way to address the measurement problem. Other research following a naturalistic approach argues that strategy is the appropriate level at which to study problem solving (Klein, 1989; Zsambok, Beach, & Klein, 1992). One advantage of studying problem solving from a strategies perspective is the recognition that strategies are flexibly applied to adjust to the constraints and requirements of a situation. Strategies are sequences of processes that are intelligently and adaptively used by problem solvers to manage the accuracy-effort trade-off in performance. A review of about 200 literature sources resulted in the identification of 66 strategies in three classes: managing information, controlling progress, and making choices (Pounds & Fallesen, 1994). These strategies were not viewed as inclusive of all possible strategies, but as a baseline tied to specific theoretical and empirical literature. Research on this baseline should indicate the sufficiency of the strategies.

Research Method

Research was designed to collect information on strategy use and perceived importance of the strategies for three problems. So far the experimental protocol has been applied to structured interviews with 32 officers. The first problem that the technique is applied to is one from the participant's own experience. In an interview the officers are asked to tell about a tactical problem that is particularly memorable, what the problem was, how the problem came to be recognized, and how it was solved. The interviewer seeks clarification as necessary while the story is reported. The participant is then asked to sort 29 strategies described on index cards into groups: not used, uncertain, used (but not important), important, very important, and most important. The 29 cards address various information processing strategies. Participants are asked to give examples of how the strategies were descriptive of how they thought. Each participant is then asked to do the same kind of sorting and description for 19 choice strategies.

Two more problems were presented to the participant. One of these was the tactical problem described in the introductory paragraph. The second was similar but places the participant in the role of an infantry company commander who has been training his company in another country and is assigned to rescue a U.S. Ambassador and his family from their captors. Participants also indicate the proportion of time that they use one of four problem solving approaches in their everyday life. The four approaches consist of short paragraphs describing

(1) a multi-attribute utility analysis (MAUA) approach,
(2) a concurrent procedural approach (Step),
(3) a recognition-primed decision approach (RPD),
(4) a dominance structuring approach (Dom).

The first two approaches follow the classical model and the last two are consistent with the naturalistic model. MAUA is an extreme of the classical model focusing on very explicit, quantitative comparisons. The Step approach describes the problem solving process of breaking the process into steps following the classical model. The Step approach is typical of instruction in the CAS3 program and many elements of the command estimate doctrine. RPD describes a process of recognition, where responses are quickly made based on the strength of cues. When recognition does not occur quickly, the problem solver identifies and explores possibilities and can accept the first solution that will work (a process of **satisficing and progressive deepening**). Dominance structuring describes a process where a candidate solution is interpreted as dominant (or is rejected) based on the pro (or con) arguments the problem solver can determine.

Table 4
Use and Importance of Problem Solving Strategies: Experimental Procedure

Participant's tactical experience	**CoCdr Rescue the Ambassador**	**Bn Cdr Enemy over bridge**	**Data Collection**
What was problem? How solved?	What was problem? How solved?	What was problem? How solved?	Interview, verbal protocol
Which strategies were used?	Which strategies were used?	Which strategies were used?	Card sort
How important was each?	How important was each?	How important was each?	Card sort
What approach?	What approach?	What approach?	MAUA, Step, RPD, Dom
Self-assessment of outcome	Self-assessment of outcome	Self-assessment of outcome	Rating scale

Three sets of measures (strategy importance, outcome ratings, and everyday approaches) along with demographic information on the participants make up the data that have been analyzed to date. Future analyses of the verbal protocol recordings will confirm the self-reported strategies and will allow a richer understanding of the blend of strategies and quality of solutions.

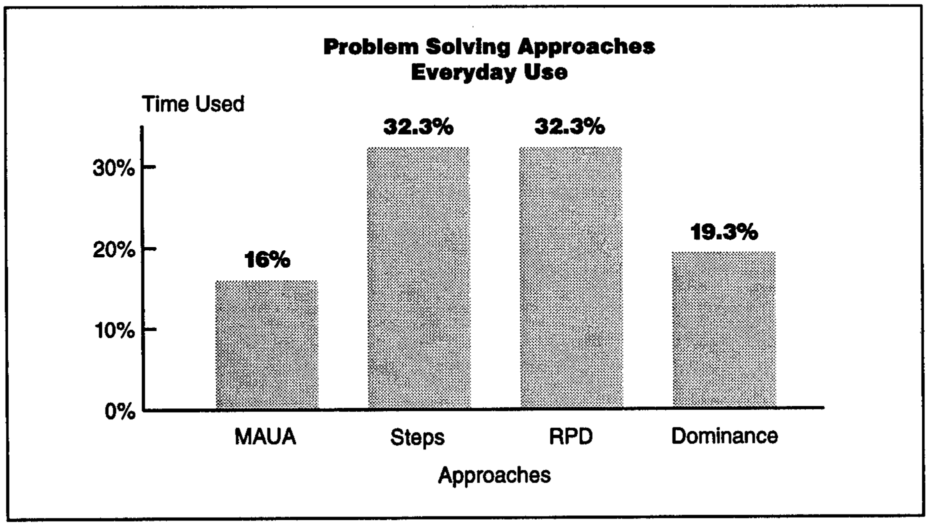
Results on Problem Solving Strategies

Effects of military schools and rank

There were significant differences among some strategies based on either military schools graduated from or by rank. One such strategy is "restating the problem" (P10). This is important in both naturalistic and classical models, but in the study Lieutenants almost never re-stated the problem (p=.005). Although the classical teachings of CGSOC recommend the "explicit identification of facts" (P21), CGSOC graduates used this strategy less than non-graduates (p=.01). The classical model is characterized by the call to "suspend judgment until all options are considered" (P29). However, this strategy was preferred by those who had less instruction in this. Lieutenants used "suspending judgment" more than Captains and Majors, and Captains and Majors used this strategy more than Lieutenant Colonels (p=.10). There are several good reasons for "reexamining options" (C19), but participants who were CAS3 graduates reexamined options less than those who had not attended CAS3 (p=.05).

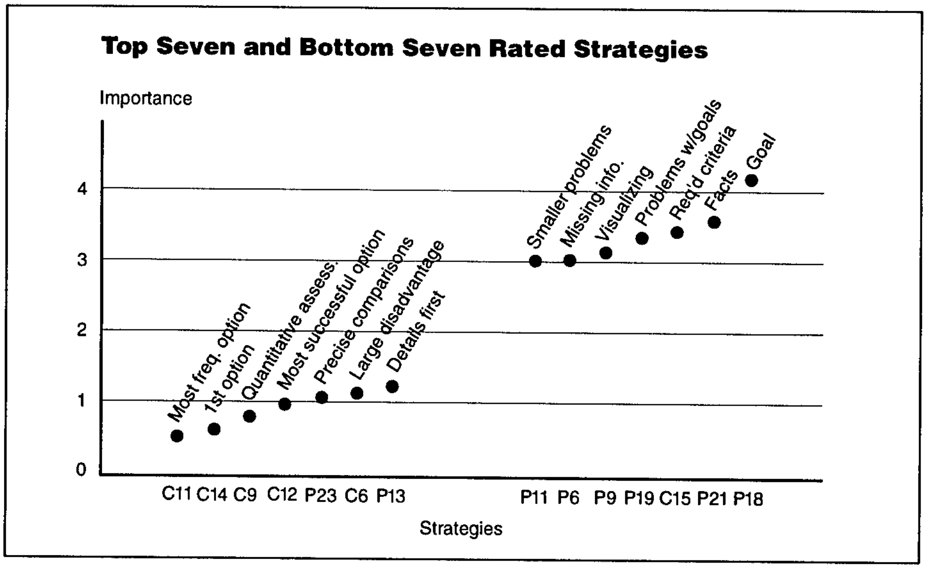
Problem solving approaches

The graph shows that preferences for the classical and naturalistic approaches are equally split between the Step and RPD approaches. It could be expected that there is some bias favoring the classical methods since they are the approaches officers should be most familiar seeing as an explicit description. It is encouraging that the ratings for the naturalistic strategies were as high as they were.



Predicting approaches with strategy subsets

Each of the 48 strategies were predicted to be positive, negative, or neutral indictors of the MAUA, Step, RPD, and Dom approaches. The strategies were ordered according to their average importance ratings. The top seven and bottom seven strategies were contrasted to determine any discernable patterns. Five of the top seven strategies are common indicators among the MAUA, Step, RPD, and Dom approaches. "Considered what information was missing" (P6) is unique to the RPD approach and "identified facts" (P21) is unique to the Step approach. The bottom 6 strategies are key, unique markers for the MAUA, RPD, or Dom approaches. Since the top strategies tended to be common across approaches and the bottom strategies tended to be unique, problem solvers appear to cross style divisions and make up more complex combinations.


Top Seven and Bottom Seven Rated Strategies

Examining the relationship more closely allows testing the adequacy of the hypothesized fit of strategies to approaches. For instance, 16 strategies were hypothesized to define the MAUA approach. Some would be positive indications of the approach (e.g., "used specific and precise comparisons" [P23], "suspended judgment" (P29), "quantitative evaluation of options" [C9]). Other strategies would counter-indicate the MAUA approach (e.g., "looked at the problem as a story" [P7], "used general and approximate comparisons [P24], "chose the option that had occurred most often" [C11]).

To test the notion that strategies could predict approaches, each strategy was considered to be a positive, negative, or neutral indicator of each of the four approaches. Valence was determined by reviewing descriptive and instructional materials. The positive and negative indicators were regressed against the theoretical approach they were to represent. The table shows squared regressions for the hypothesized strategies and for strategies determined by Mallows' select criterion (1973) (when the number of predictors approaches an estimate of the total squared error). No "best" regression is reported for MAUA because the number of predictors did not approach the error estimate until the full set of predictors was used. The hypothesized sets of strategies had modest prediction levels for the approaches. Smaller sets of strategies were found for the Step, RPD, and Dom approaches using Mallows' criterion.

Table 5
Hypothesized and "Best" Regression Models

	Hypothesized Set of Strategies		**Strategies Determined by Mallows' Criterion**		**Model Comparison**		
Approach	Shared Variance R^2	No. of Strategies	Shared Variance R^2	No. of Strategies	DF	F Ratio	Sign. Level
MAUA	.22	16					
Step	.36	20	.61	11	61,70	1.864	.0061
RPD	.51	27	.45	6	54,75	1.231	.2011

Only the alternate predictor set for the Step approach was significantly better than the hypothesized set. Comparison of the particular strategies that went into the hypothesized and Mallows-determined regression models revealed that only three of the strategies were common, and one of these had a negative instead of positive weight (C17). Four of the hypothesized strategies that were considered positive indicators actually contributed negatively to the prediction. The Mallows-determined model for the Step approach consisted of the following predictor strategies with corresponding parameter weights.

Table 6
"Best" Model of Individual Strategies Regressed on Step Approach

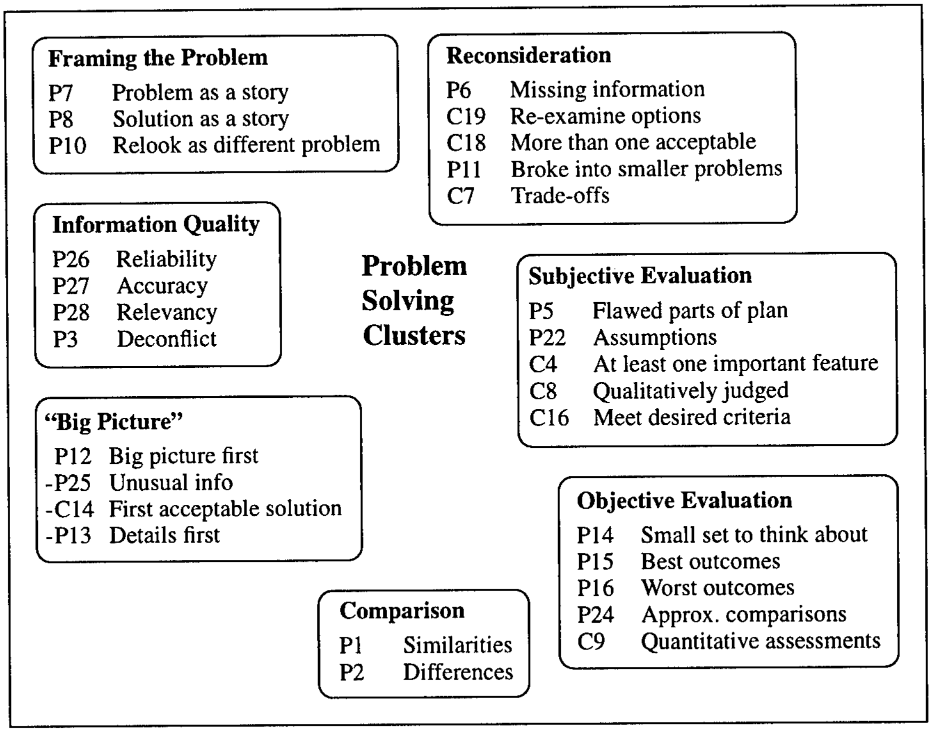
Positive Indicators			Negative Indicators		
Code	Strategy	Weight	Code	Strategy	Weight
P5	Identified flawed parts of plan	3.61	C1	Kept things to think about small	-3.23
C3	Used standards set by others	3.30	P17	Considered various perspectives	-2.59
P18	Identified specific goal	3.08	P15	Imagined best outcomes	-2.44
C19	Reexamined acceptable options	2.50	C6	Considered large disadvantages	-2.30
P26	Information reliability	2.43	C17	Eliminate some options	-2.26
P23	Used specific comparisons	2.12			

Consideration of these predictor strategies shows several unexpected relationships. For example, strategy P23 should be a positive indicator of MAUA and negative indicator of RPD but a neutral indicator for the Step approach. Other strategies predicted to be neutral indicators of the Step approach included the following that should be characteristic of the other approaches. Strategy C1 should be a positive indicator of RPD and Dom and a negative indicator of MAUA. Strategy P17 should be a positive indicator of RPD. Strategy C6 should be a positive indicator of Dom. Strategy P15 should be a negative indicator of RPD.

The Step model is not perceived by actual problem solvers to include the same strategies as included in instructional and doctrinal guidance but to include strategies that conceptually are indicators of the naturalistic approaches. The hypothesized strategies for RPD and Dom were modest but adequate predictors, no different from the unbiased regressions determined by Mallows criterion.

Cluster Analysis

The importance ratings were used in factor and cluster analysis to identify strategies that went together. Because of high partial correlations, 20 of the 48 strategies were eliminated based on Kaiser's measure of sampling adequacy (Cerny & Kaiser, 1977) and the percentage of everyday use of Step and RPD approaches were added. Principal components were computed for these 30 measures, they were rotated using an oblique procedure, factor scores computed, and these scores were cluster analyzed (see SAS, 1988). Seven clusters explained 49% of the variance. The percentage of usage of the strategies averaged for each cluster were: subjective evaluation, 77%; "big picture," 72%; reconsideration, 72%; information quality, 69%; comparison with experience, 62%; objective evaluation, 59%; and framing the problem, 39%.



Problem Solving Clusters

Framing the Problem

P7	Problem as a story
P8	Solution as a story
P10	Relook as different problem

Reconsideration

P6	Missing information
C19	Re-examine options
C18	More than one acceptable
P11	Broke into smaller problems
C7	Trade-offs

Information Quality

P26	Reliability
P27	Accuracy
P28	Relevancy
P3	Deconflict

Subjective Evaluation

P5	Flawed parts of plan
P22	Assumptions
C4	At least one important feature
C8	Qualitatively judged
C16	Meet desired criteria

"Big Picture"

P12	Big picture first
-P25	Unusual info
-C14	First acceptable solution
-P13	Details first

Objective Evaluation

P14	Small set to think about
P15	Best outcomes
P16	Worst outcomes
P24	Approx. comparisons
C9	Quantitative assessments

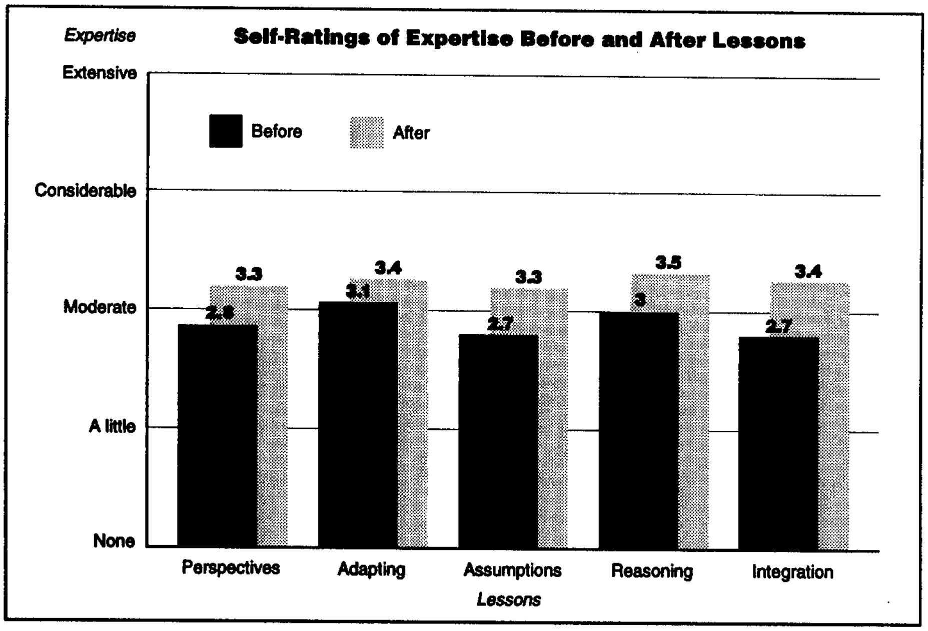
Comparison

P1	Similarities
P2	Differences

Utilization of Research

In a study commissioned by General Franks while he was TRADOC Commander, ARI reported that guidance that Army leaders receive on problem solving is primarily based on the classical model (Halpin, in preparation). The current research findings point out that the strategies that officers find important are more varied than what is taught based on the classical model. Many of the alternate strategies that are potential improvements are not viewed as important, suggesting a need for additional instruction. For example, four strategies that are related to characteristics of expert behavior were found to be in the bottom quartile of use (P10-Restating the problem, P17-Taking multiple perspectives, P25-Considering unusual information, and P28-Considering information relevancy).

GEN Franks and LTG Miller (Commandant of CGSC) requested that an experimental course be developed to address additional ways of solving problems. The course, called practical thinking, was developed and applied to the Mobile Strike Force element of Prairie Warrior '95. Several lessons making up 16 hours of instruction were presented to 73 students of the Battle Command course in CGSOC. Practical thinking was one of many components of the Battle Command course. The other parts of the course included an alternate division staff con-cept (where officers are assigned to general-purpose positions); use of 2010 weapons and systems technology; development of doctrine, tactics, techniques, and procedures for 2010 capabilities; information systems capabilities with a command decision support system (Phoenix); a General Officer to lead the class and student-manned division headquarters; and frequent use of whole-staff simulation training. Within this very dynamic and challenging setting, the practical thinking instruction resulted in an average gain of .52 points (on a 5-point scale) in self-reported expertise. Eighty percent of the students responding to the final course evaluation indicated that they felt the instruction should be offered to future CGSOC students. Some felt that it was equally valuable at higher and lower levels in an officer's formal development. A few students felt that this was the most important part of the Battle Command course.



BG Geoffrey Miller, Commander of the Mobile Strike Force Division, recommended to LTG Miller that the practical thinking instruction be included in subsequent Battle Command courses. The Leadership Instructional Department has already incorporated several aspects of the practical thinking instruction into their core and elective courses. Requests for materials on practical thinking have come from the Army War College, the USDA Forest Service, the Michigan State Police, and the Los Angeles County Fire Department.

Summary

Fundamental research into strategies for tactical problem solving is long overdue. Instructional guidance based on classical models of human behavior does not provide an adequate basis for improving problem solving. Instructional concepts with a broader focus seem to have promise. These concepts attempt to increase a problem solver's tendency to reflect about their style of thinking and to provide some basic tools that can be molded to fit the situational demands of future problems. Naturalistic approaches have started to make an impact on doctrine and instruction.

The ARI problem solving strategy research will lay the groundwork for a more considered set of recommendations for problem solving procedures. The research has already established a catalog of problem solving strategies that can be explicitly measured. Tactical leaders are being surveyed on the use and importance of strategies in standard test problems and actual problems they have experienced.

In future research, the strategies used by novices and experts can be distinguished so instructional plans can be determined more deliberately. With standard definitions of problem solving strategies, leaders can be trained to be more reflective about the processes that acually occur during problem solving. Such are the anticipated benefits of this research to better understand actual problem solving behavior.

References

Cerny, B. A., & Kaiser, H. F. (1977). A study of measure of sampling adequacy for factor-analytic correlation matrics. *Multivariate Behavioral Research,* 12, 43-47.

Essens, P., Fallesen, J., McCann, C., Cannon-Bowers, J., & Dorfel, G. (1995). *COADE —A framework for cognitive analysis, design, and evaluation.* Technical Report AC/243(Panel 8)TR/17. NATO Defence Research Group.

Fallesen, J. J. (1995). Decision matrices and time in tactical course of action analysis. *Military Psychology,* 7(1), 39-51.

Fallesen, J. J. (1993). *Overview of Army tactical planning performance research.* (ARI Technical Report 984). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A273 273)

Fallesen, J. J., Lussier, J. W., & Michel, R. R. (1992). *Tactical command and control process.* (ARI Research Product 92-06). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A255 036)

Klein, G. A. (1989). Strategies of decision making. *Military Review,* 56-64.

Mallows, C. L. (1973). Some comments on Cp. *Technometrics,* 15, 661-675.

Pounds, J. F., & Fallesen, J. J. (1994). *Understanding problem solving strategies.* (ARI Technical Report 1020). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A290 350)

SAS Institute Inc. (1988). *SAS/STAT™ User's Guide, Release 6.03 Edition.* Cary, NC: SAS Institute Inc.

Zsambok, C. E., Beach, L. R., & Klein, G. A. (1992). *A literature review of analytical and naturalistic decision making* (Technical Report No. N66001-90-C-6023). San Diego, CA: Naval Command, Control and Ocean Surveillance Center.

Small Unit Dynamics: Leadership, Cohesion and Motivation

Guy L. Siebold, Ph.D.
U.S. Army Research Institute
Organization and Personnel Resources Research Unit,
Alexandria, VA

ARI has been conducting advanced research to refine scientific knowledge about small combat units and their performance as well as develop pertinent policy information and products concerning them. Research has focused on advancing theory, improving measurement, and building models relating key "small group" variables such as cohesiveness, climate, leadership, and motivation. Results from questionnaires, for example, have been extremely predictive of later unit performance (1) on tactical field exercises, (2) on Operational Readiness Evaluations, and (3) at the Joint Readiness Training Center. The Platoon Cohesion Index questionnaire or its derivatives has been used successfully for units in different Army branches and in the Active and Reserve Components. The Index has been translated and used with similar results in Israel (Hebrew) and Canada (French and English). [Translation into Spanish is forthcoming.] Short, self-administered questionnaires have been made available for field use. The strong measures, which are highly predictive of unit performance, have been provided to other agencies for use as surrogate performance criteria. The measures have also been used as criteria to assess the impact of incremental changes in training resources. Policy relevant analyses have found, for example, (1) that the racial/ethnic group mix (heterogeneity) in a platoon is unrelated to its cohesion, level of motivation, or performance, (2) that occupants of vestigal (nominal) positions during peacekeeping deployments substantially decrease their level of mission motivation, and (3) that Reserve Component units engaged in peacekeeping look and change very much like Active Component units in terms of leadership, cohesion, and motivation. These research efforts have been documented by Dr. Siebold in over 30 reports, chapters, articles, and papers.

Training Transfer: An Empirical Comparison of Two Training Development Approaches

Dorothy L. Finley, Ph.D.

U.S. Army Research Institute

Armored Forces Research Unit

Fort Knox, KY

The High Transfer Training (HITT) approach to developing a program of instruction (POI) is an extension of the Systems Approach to Training (SAT). SAT is the current conventional means of training development in the U.S. Army. The HITT extension supports analysis and design of training programs for jobs which require performing actions on several differing objects or object configurations. If certain commonalities are found to exist between at least some objects, then analyses can be performed to enable design of HITT strategies into the POI. These analysis and design efforts lengthen the development process somewhat. HITT implementation may be more or less expensive, depending on several factors. Summative evaluations had already established that a HITT-developed POI produced students who both met course standards and were able to transfer their training to new but similar equipments. This experiment went a step further to determine whether the HITT POI provided any training transfer value beyond that afforded by a conventional POI. Students nearing completion of a HITT developed POI were compared to students nearing completion of a conventionally developed POI. The POIs provided Advanced Individual Training for different, but related, job specialties. The jobs were judged to encompass sufficient similarities to allow POI comparisons. Students from the two POIs performed operations and maintenance tasks on an equipment item for which their branch, Signal, was also the proponent. It was, however, an item for which neither group was responsible and on which neither group had received specific training. Findings supported the hypothesis that the HITT POI provided significantly greater transfer training value. Perhaps more important, however, is the possible inference that the additional value was gained through differences in the approaches to developing and designing the POIs. This is evidence, therefore, that—when conditions are appropriate—the value gained may be worth additional costs.

Deriving Lessons Learned From the U.S. Army Combat Training Centers: An Opposing Force Perspective

Robert F. Holz, Ph.D.
U.S. Army Research Institute
Advanced Training Methods Research Unit
Alexandria, VA

The training carried out at the Combat Training Centers (CTCs) utilizes both live fire and force-on-force exercises. In the latter, opposing force (OPFOR) soldiers portray a credible enemy utilizing appropriate doctrine, equipment, and organization to conduct realistic, two-sided, free-play exercises designed to stress blue force units (BLUFOR) to the maximum.

Training at the CTCs is designed to provide BLUFOR with doctrinally based feedback through the mechanisms of After Action Reviews (AARs) and Take Home Packages (THPs). Virtually all feedback is provided by observer controllers (O/Cs) who have previously commanded the same echelon they are observing and have received special training that enables them to perform this essential function.

The OPFOR play only a minor role in contributing to the feedback process because of the heavy demands placed on them to fight the BLUFOR and to maintain their own combat readiness. However, they are uniquely positioned to provide important lessons learned regarding BLUFOR strengths and weaknesses amenable to solutions based on doctrine, training, organization, leader development, and soldier (DTMOLS) systems.

The information secured in this study was derived from interviews carried out with a sample of OPFOR and O/C personnel at the Combat Maneuver Training Center.

Analyses of the interviews revealed that O/Cs tended to focus their assessments of BLUFOR in terms of broadly based doctrinal solutions. On the other hand, members of the OPFOR tended to emphasize relatively specific tactics, techniques, and procedures (TTPs).

While time demands preclude extensive input by the OPFOR during AARs, the OPFOR could be called on to provide more specific feedback regarding BLUFOR during breaks between rotations. Such feedback could be used for developing lessons learned along with input from O/Cs.

Device-Based Prediction of Tank Gunnery Performance

Joseph D. Hagman, Ph.D.
U.S. Army Research Institute
Reserve Component Training Research Unit
Boise, ID

To determine the relationship between device- and tank-based gunnery performance, two groups of 29 Armor crews were tested on the Conduct-of-Fire Trainer (COFT) 1 day before undergoing live-fire Table VIII evaluation. A significant ($p < .05$) positive relationship between COFT and Table VIII scores was found for Group 1 (i.e., the normative group) and confirmed for Group 2 (i.e., the cross-validation group), with the Group 1 predictive model accounting for over half the variance in the live-fire scores of both groups. A practical tool was then developed from pooled data to help U.S. Army National Guard unit trainers accurately predict the probability of successful first-run Table VIII qualification for individual tank crews. Additional research needed to maximize the payoff from prediction tool usage is also discussed.

Measuring Propensity of African-American Youth to Join the Military

Joel M. Savell, Ph.D.
U.S. Army Research Institute
Organization and Personnel Resources Research Unit,
Alexandria, VA

Trend data from the Defense Manpower Data Center's Youth Attitude Tracking Survey (YATS) suggest that, over the past several years, there has been a decline in "propensity" to join the military—particularly for male African-American youth. There is some uncertainty as to whether this decline is real and, by extension, whether the military services will have difficulty meeting their recruiting goals. Whatever the answer to this question, however, the African-American youth population is an important one for military planners, and the Army needs to know more about it than it presently does. It needs to know what the relevant variables are, and it needs to know how to measure them.

This research is investigating reference group influences on the attitudes and values of African-American youth. The question here is whether some of these reference groups (or reference persons) are more important in this respect than others. Our first step, carried out this year, was to develop a nonreactive procedure that could be used experimentally to demonstrate attitudinal influence on these youth by a particular reference group (the subject's close friends). We interviewed 143 students regarding the importance—to themselves and also to their close friends—of selected attitude objects (e.g., having a job that most people look up to and respect). Half the interviewees gave ratings for themselves first and then estimated ratings for their friends, while the rest gave these ratings in the reverse order. As hypothesized, students who gave their ratings after estimating ratings for their close friends—and presumably thinking about these ratings—gave self-ratings that were closer to the ratings they had estimated for their friends than was the case for students who gave ratings for themselves first ($p<.001$).

The second step, which is projected for next year, is to adapt the procedure for use in comparing two or more reference persons or groups, e.g., parents and friends.

Psychophysics of Perceptions With a Virtual Reality Helmet Display

Robert H. Wright, Ph.D.
U.S. Army Research Institute
Rotary-Wing Aviation Research Unit
Fort Rucker, AL

The ARI STRATA helicopter training research simulator was used to define perceptual performance with its computer-generated high resolution 65 by 125 degree helmet display.

In one study a magnitude production was used in which subjects used joysticks to set requested forward or lateral distances, heights, or speeds. Median forward distance and speed perceptions were 41% of actual, lateral distance 50%, and height perceptions 72%. Relative perceptions were more accurate than absolute, and increasing distance/speed perceptions more accurate than decreasing. High visual database detail with familiar objects had only a slight positive effect over just ground texture. Perceptual accuracy decreased as the offset of the line of eyepoint motion from the visual reference point increased.

The second study determined thresholds for the fore-aft, lateral, and vertical motion perceptions that are required in hovering a helicopter. Subject's perceptual response time and vehicle drift for 43 different visual scene conditions were obtained. These were measured during 20 repeated 7-second presentations of 1, 2, or 3 orthogonal winds that varied in sign and speed. Median thresholds ranged from 20 to 100 cm. The standard gradual change antialiased ground texture provided best overall motion perception. Height, fog, and shadows had limited effect. Vertical motions were perceived best, and the visual scene factors had little effect on them. Upward motion was perceived better than downward motion. Rearward motion was perceived better than forward motion, with larger differences for more distant trees. The better motion parallax cues of multiple rings of trees improved mainly lateral motion perceptions. Fore-aft and lateral motion perception decreased with increasing tree ring distance. Vertical motion had substantial interaction effects on the perception of fore-aft and lateral motions.

Study of Training Aids, Devices, Simulators, and Simulations (TADSS)

Robert H. Sulzen, Ph.D.
U.S. Army Research Institute
National Training Center Element
Fort Irwin, CA

The U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) was asked by the Training and Doctrine Command (TRADOC) to conduct a study of Training Aids, Devices, Simulators, and Simulations (TADSS) and to recommend a system for routinely gathering TADSS utilization data. The objectives for this study were to: (1) identify how the available TADSS are integrated into training programs, (2) evaluate user perceptions of TADSS, and (3) provide recommendations for a procedure to periodically gather this information.

The methodology consisted of administering structured interviews at each of eight posts with selected personnel. Twelve interview guides were developed. TADSS are utilized differently by the different branches, that is, infantry and armor are the primary users of the Unit Conduct of Fire Trainer (UCOFT). Constructive Simulation is widely used by companies and battalions and not often by platoons. UCOFT and Weaponeer are the simulators most often used. Virtual Simulation is limited to the use of the combat maneuver arms where it is available. Live Simulation with Multiple Integrated Laser Engagement System (MILES) is the TADSS most often mentioned, and as a system has the greatest differences in the perception of use between leaders and their subordinates. TADSS information should be collected on a periodic basis, and a modified Standard Army Training System (SATS) might be adapted for this process.

Economic Life Course Analysis of Peacekeeping in the Sinai

Hyder Lakhani, Ph.D.

U.S. Army Research Institute

Organization and Personnel Resources Research Unit

Alexandria, VA

We remain prepared to support traditional peacekeeping operations. ...Reserve component elements will take on increased responsibility for participation in and supporting peacekeeping missions. Joint Chiefs of Staff. National Military Strategy of the USA, 1995, p. 9.

The relatively greater downsizing of the Active Component (AC) compared with that of the Reserve Component (RC) has increased the importance of the RC for peacekeeping and other missions. The 28th rotation of the Multinational Force and Observers peacekeeping mission in the Sinai consisted of 80% RC and 20% AC. The RC soldiers face an economic tradeoff in volunteering for peacekeeping because they lose their civilian earnings but receive regular military earnings instead of the token drill pay. The net economic gain/loss can increase/decrease their Army career commitment and volunteering for future missions. Therefore, this project collected survey data on financial gains/losses and for other demographic and organizational variables during training and deployment phases. The data revealed that, in general, the RC gained and the AC lost during the training phase. A procedure of multiple regressions was used to predict intentions to stay up to 20 or more creditable years. The results revealed that the soldiers were more likely to stay with an increase in financial gain (basic monthly pay), length of service, and home ownership. Also, the RC soldiers were more likely to stay (and less likely to quit) than AC. Therefore, policy makers should continue future deployment of RC for such missions.

Future research will analyze the data collected during deployment phase and longitudinal data to be collected during three postdeployment phases. These analyses will help discern trends in economic life courses of the volunteers and the correlation between intentions and behavior. The results for these soldiers will be compared with the analysis of data collected for a control group of soldiers in the 29th LID who did not volunteer to go to the Sinai.

Racial Attitudes of White Veterans Toward Blacks

George H. Lawrence, Ph.D.
U.S. Army Research Institute
Organization and Personnel Resources Research Unit
Alexandria, VA

Data from the General Social Survey were analyzed to test the hypotheses that (1) white veterans would express more positive attitudes toward blacks than nonveterans and (2) attitudes of white veterans whose military service was in an equal opportunity (post 1975) military would be more positive toward blacks than veterans whose service occurred earlier. Racial attitudes of white veterans and nonveterans differed relatively little after controlling for effects of age, education, and year of survey response. While veterans were slightly more likely to be against special governmental obligation or assistance to black citizens, they were also slightly more apt to say they had recently entertained a black in their homes for dinner. For a limited set of variables, white veterans who served after 1975 expressed slightly more negative attitudes toward blacks than did white nonveteran controls. Methodological constraints are discussed and alternative interpretations of the data are offered. The conclusion is suggested that there is no simple way to translate the relative racial harmony that exists within the military to civilian contexts.

Assessment of User Reactions to the Multi-Service Distributed Training Testbed (MDT2)

Angelo Mirabella, Ph.D.
U.S. Army Research Institute
Advanced Training Methods Research Unit
Alexandria, VA

In May 1994 and February 1995, ARI joined with the Air Force, Navy, and Marine Corps to developmentally test and demonstrate distributed training methodology. In a week of close air support exercises, we tested training and evaluation methods specially designed for use with multiService distributed interactive simulation (DIS). The goal was to create a model to define training objectives, convert those into exercise scenarios, conduct the training, and provide afteraction reviews. As part of the total system assessment, ARI surveyed trainee and observer/controller reactions. Three questions were addressed—

What value does multiService distributed training add to the instructional pipelines of the Services?

How well were training objectives satisfied by the simulated combat exercises?

How useful were various feedback techniques (e.g., rapid plan-view replay of exercises, video teleconferencing of AARs, and three-dimensional stealth displays)?

Data analysis is still in process, but highlights of preliminary results will be presented.

Span of Command and Control: Implications of New Research for Designing Organizations

Richard E. Christ, Ph.D.

U.S. Army Research Institute

Fort Leavenworth Research Unit

Changes in operating environments are having a significant impact on all types of organizations. The impact is most dramatic when the changes are associated with downsizing in resources even as new demands are placed on the organization. One important aspect in designing any organization is the concept of Span of Command and Control (SOCC). However, there is little if any reliable data that can be used to specify the precise nature of this relationship or the factors which moderate the form of the relationship.

During September 1993 to March 1994, the project team interviewed 11 U.S. Army General officers regarding operations other than war and 44 officers from Captain to Lieutenant General regarding war fighting operations. The interviews were structured around seven factors proposed as affecting the SOCC: task characteristics, organizational structure, complexity of environment, history or unit continuity, technology, individual characteristics, and external organizations. The data collected consisted of the comments made during the interviews, the results of a content analyses of those comments, and, for war fighting operations only, ratings on the impact of each factor on the difficulty of command and control. Both sets of data were examined as a function of level in the organizational hierarchy or echelon, and type of organization. The results show an interacting effect of factor, echelon, and type of organization on the difficulty of command and control. This presentation summarizes these results and presents conclusions and recommendations for organizing Army units based on the results.

Metacognitive and Social Processes in Team Training

Ray S. Perez, Ph.D.

U.S. Army Research Institute

Advanced Training Methods Research Unit

Alexandria, VA

Combat units such as platoons, squads, crews, and team are the primary engagement elements of the Army. The training of these units has been a key concern of the Army leadership as the types of mission and battlefield conditions become more diverse and complex. However, unit training has not always been conducted with systematical rigor nor has there been a empirical basis to guide the development of unit training exercises.

To support these training requirements, ARI has been engaged in research to develop a methodology to generate strategies that produce units that are flexible and adaptable. This methodology combines research from training areas such as collective problem solving, metacognition processes, and cooperative learning. This research has provided evidence of the effectiveness of these training strategies and variables such as coordination, cooperation, communication, and metacogntion (see Slavin, 1989, 1990; Johnson & Johnson, 1989a; Brown, 1979; & Sternberg, 1984) on team problem solving. Sharan (1980) and Meloth (1992) have argued that the most important result of this research is the role of metacognitive strategies and social skills used by subjects during learning to problem solve in a group/team.

The theory used for studying the role of metacognitive behavior and social processes in collective problem solving has been proposed by Sternberg (1984). In his Triachric Theory of intelligence he carefully identified the role of metacognitive components in the problem solving process. This view of intelligence suggests that the use of metacognitive components during problem solving characterize intelligent behavior. Within this framework teams who exhibit the use of these metacognitive components during problem solving are viewed as intelligent. A series of experiments have been conducted to determine effects of social skills and metacognitive processes on team performance. Results indicate that combining social skills with metacognitive processes facilitates team problem solving that is superior to those trained by either social skills or metacognitive processes.