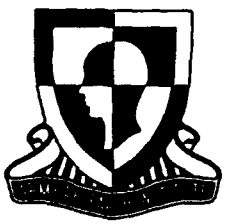
Technical Report 1006

Training Research with Distributed Interactive Simulations: Lessons Learned from Simulation Networking

John A. Boldovici and David W. Bessemer
U.S. Army Research Instutute

August 1994



United States Army Research Institute for the Behavioral and Social Sciences

U.S. ARMY RESEARCH INSTITUTE
FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency Under the Jurisdiction of the Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON
Director

Technical review by

Dwight J. Goehring
Ronald C. Hofer, STRICOM
Don Johnson

NOTICES

DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-POX, 5001 Eisenhower Ave., Alexandria, Virginia 22333-5600.

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	1994, August	Final Jan 93 - Sep 93

4. TITLE AND SUBTITLE
Training Research with Distributed Interactive Simulations: Lessons Learned from Simulation Networking

5. FUNDING NUMBERS
63007A
795
2114
H1

6. AUTHOR(S)
Boldovici, John A.; and Bessemer, David W.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army Research Institute for the Behavioral and Social Sciences
ATTN: PERI-RF
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

8. PERFORMING ORGANIZATION REPORT NUMBER
ARI Technical Report 1006

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
--

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
--

11. SUPPLEMENTARY NOTES
--

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE
--

13. ABSTRACT *(Maximum 200 words)*

Empirical and analytic evaluations of Simulation Networking (SIMNET) were reviewed to derive recommendations for planning evaluations of the Close Combat Tactical Trainer (CCTT). Lessons learned from SIMNET evaluations are as follows: (1) One-shot empirical evaluations of the kind performed to meet acquisition, test, and evaluation regulations are costly and unlikely to meet CCTT evaluation objectives; (2) analytic evaluations of SIMNET produced low-cost information that can be applied to improving CCTT design and use and in budget justifications; and (3) empirical evaluation alternatives to past methods should be considered to support CCTT evaluation objectives that pertain (a) to establishing the relation between CCTT training and soldier performance in the field and (b) to complying with acquisition, test, and evaluation regulations.

Evaluation alternatives were presented for CCTT, with discussions of the advantages and disadvantages of each. The evaluation alternatives included in-device learning experiments, quasi-transfer experiments, correlation of scores achieved in SIMNET or CCTT training with scores obtained during rotations at Combat

(Continued)

14. SUBJECT TERMS		15. NUMBER OF PAGES
Distributed Interactive Simulation	SIMNET	66
Empirical evaluation	CCTT	16. PRICE CODE
Total Quality Management	Analytic evaluation	--

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

NSN 7540-01-280-5500

Standard Form 298 (Rev 2-89)
Prescribed by ANSI Std Z39-18
298-102

13. ABSTRACT (Continued)

Training Centers, efficient experimental designs (randomized block designs, repeated-measure Latin squares, and analyses of covariance), quasi-experimental designs, improved methods for documenting training, and analytic evaluations.

Recommendations included (1) Evaluations should address how the CCTT complements or supplements existing training alternatives to support and implement Combined Arms Training Strategy while remaining within contemporary and future budgetary limitations; (2) CCTT evaluation should be a part of a larger program of Total Quality Management (TQM) applied to the Army training system and directed toward continuous improvement in training; and (3) the CCTT evaluation process should be incorporated as a continuous part of the TQM process.

Technical Report 1006

Training Research with Distributed Interactive Simulations: Lessons Learned from Simulation Networking

John A. Boldivici and David W. Bessemer
U.S. Army Research Institute

Simulator Systems Research Unit
Stephen L. Goldberg, Chief

Training Systems Research Division
Jack H. Hiller, Director

U.S. Army Research Institute for the Behavioral and Social Sciences
5001 Eisenhower Avenue, Alexandria, Virginia 22333-5600

Office, Deputy Chief of Staff for Personnel
Department of the Army

August 1994

Army Project Number
2Q263007A795

Training Simulation

FOREWORD

The Project Manager for Combined Arms Tactical Training (PM CATT), Colonel James Shiflett, requested the assistance of the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) in recommending long-term evaluation methods for the Close Combat Tactical Trainer (CCTT). In response to that request, ARI's David W. Bessemer and John A. Boldovici reviewed reports of training-effectiveness research with Simulation Networking (SIMNET) to derive lessons learned for application to CCTT. Their reviews led them to conclude that one-shot empirical evaluations of the kind performed for SIMNET were unlikely to meet CCTT evaluation objectives that pertain (1) to establishing the relation between CCTT training and soldier performance in the field, and (2) to complying with acquisition, test, and evaluation regulations. Meeting those objectives will, in the authors' view, require changes in our thinking about the role of evaluation in device development. A Total Quality Management (TQM) orientation is advocated, in which CCTT evaluation becomes an institutionalized and continuous component of the TQM process for Army training. The authors also advocate establishment of training-data archives for SIMNET and CCTT, compatible with archives for Combat Training Center data. Evaluation alternatives are presented with discussions of the advantages and disadvantages of each.

The work described in this report was performed under the PM CATT-ARI Memorandum of Understanding titled "Training Research Support of Combined Arms Tactical Trainer Development." Drafts of the report were supplied to PM CATT, to government and contractor personnel charged with CCTT development, and to evaluators from the Operational Test and Evaluation Command and the Test and Experimentation Command.

EDGAR M. JOHNSON
Director

TRAINING RESEARCH WITH DISTRIBUTED INTERACTIVE SIMULATIONS: LESSONS LEARNED FROM SIMULATION NETWORKING (SIMNET)

EXECUTIVE SUMMARY

Requirement:

This report derives lessons learned from training-effectiveness research with Simulation Networking (SIMNET) for application in planning evaluations of the Close Combat Tactical Trainer (CCTT).

Procedure:

Reports of training-effectiveness research with SIMNET were reviewed, and assessments were made about the extent to which similar research might meet evaluation objectives for CCTT. The authors derived lessons learned from the SIMNET research and used the lessons learned as bases for suggesting alternative methods for evaluating CCTT.

Findings:

The lessons learned from the review of training-effectiveness research with SIMNET are as follows: (1) One-shot empirical evaluations of the kind performed to meet acquisition, test, and evaluation regulations for SIMNET are costly and are unlikely to meet CCTT evaluation objectives; (2) analytic evaluations of SIMNET produced low-cost information that can be applied to improving CCTT design and use and that can be used in budget justifications; and (3) empirical evaluation alternatives to past methods should be considered to support CCTT evaluation objectives that pertain (a) to establishing the relation between CCTT training and soldier performance in the field, and (b) to complying with acquisition, test, and evaluation regulations.

Evaluation alternatives for CCTT were presented with discussions of the advantages and disadvantages of each. The evaluation alternatives included in-device learning experiments, quasi-transfer experiments, correlation of scores achieved in SIMNET or CCTT training with scores obtained during rotations at Combat Training Centers, efficient experimental designs (randomized blocks, repeated-measure Latin squares, and analyses of covariance) quasi-experimental designs, improved methods for documenting training, and analytic evaluations.

Recommendations include (1) Evaluations should address how the CCTT complements or supplements existing training alternatives to support and implement Combined Arms Training Strategy while remaining within contemporary and future budgetary limitations; (2) CCTT evaluation should be a part of a larger program of Total Quality Management (TQM) applied to the Army training system and directed toward continuous improvement in training; (3) the CCTT evaluation process should be incorporated as a continuous part of the TQM process.

Utilization of Findings:

The findings in this report can be used to design and conduct research with CCTT that will meet the following four evaluation objectives: (1) To permit valid inferences about the relation and contribution of CCTT training to unit performance in Combat Training Centers and in other field exercises; (2) to justify expenditures and support budget requests for continuing development of CCTT and Distributed Interactive Simulation (DIS); (3) to meet the intent of acquisition, test, and evaluation regulations; and (4) to yield recommendations for increasing CCTT and DIS training capabilities.

TRAINING RESEARCH WITH DISTRIBUTED INTERACTIVE SIMULATIONS:
LESSONS LEARNED FROM SIMULATION NETWORKING

CONTENTS

CONTENTS (Continued)

LIST OF TABLES

LIST OF FIGURES

TRAINING RESEARCH WITH DISTRIBUTED INTERACTIVE SIMULATIONS: LESSONS LEARNED FROM SIMULATION NETWORKING[1]

Introduction

Maintaining readiness in times of declining opportunities for field training demands that simulators and simulations be used to accomplish major portions of the U.S. Army's Combined Arms Training Strategy (CATS). The core of the Army's simulators and simulations for maneuver training will be the Close Combat Tactical Trainer (CCTT) and related Distributed Interactive Simulation (DIS) components. CCTT is an in-progress step between the prototype Simulation Networking (SIMNET) system and future DIS evolutions.

The importance of CCTT and DIS training to achieving CATS objectives requires that steps be taken to ensure that CCTT development yields products that will be used to advantage by Army schools and units. This requirement applies to evaluation planning for CCTT. Lessons learned from SIMNET evaluations should be applied to CCTT evaluations for purposes of increasing the efficiency of evaluations and the utility of their results.

Rationale

Empirical transfer-of-training evaluations of the kind performed to meet acquisition, test, and evaluation regulations (AR 70-1, TRADOC Regs 71-9 and 350-4) yielded results that permit no valid inferences about transfer of SIMNET training to soldiers' performance in field settings. Those empirical evaluations also yielded little diagnostic information that can be used to increase the potential training effectiveness of SIMNET or CCTT. These consistent evaluation failures with SIMNET make it appropriate to examine their origins and to explore alternatives.

---

[1]We thank Barbara A. Black, Frederick J. Brown, Stanley Bolin, G. Gary Boycan, Billy Burnside, Dwight J. Goehring, Stephen L. Goldberg, Jack H. Hiller, Ronald C. Hofer, Don Johnson, Edgar A. Johnson, Larry L. Meliza, John E. Morrison, William C. Osborn, Mark Palmissano, Michael J. Singer, and James E. Shiflett for reviewing drafts of this report.

Purpose

The purposes of this report are to

1. Identify evaluation objectives for the CCTT.

2. Examine the advantages and disadvantages of empirical and analytic evaluations of SIMNET to derive recommendations for meeting CCTT evaluation objectives.

3. Describe the reasons that CCTT evaluation objectives cannot be met by one-shot, empirical, transfer-of-training evaluations of the kind performed for SIMNET in response to acquisition, test, and evaluation regulations.

4. Suggest evaluation methods that are less expensive and more likely to meet CCTT evaluation objectives than were the methods used with SIMNET.

Evaluation Objectives

Evaluations of CCTT should have objectives whose achievement will

1. Permit valid inferences about the relation and contribution of CCTT training to unit performance in Combat Training Centers (CTCs) and in other field exercises.

2. Justify expenditures and support budget requests for continuing development of CCTT and DIS.

3. Meet the intent of acquisition, test, and evaluation regulations.

4. Yield recommendations for increasing CCTT and DIS training capabilities.

Two kinds of evaluations are required to meet the four evaluation objectives listed above. The two kinds of evaluations are empirical and analytic.

Empirical Evaluations

Empirical evaluations use statistical analyses of scores that reflect soldiers' performance in training and on tests. The statistical analyses are done to determine whether and the extent to which alternative training regimens produced performance differences between compared groups greater than

performance differences that would be expected by chance. (Discussions of basic statistical concepts and the need for statistics are in Appendix A.)

Empirical evaluations must be carefully structured in advance to permit the use of proper statistical methods and to enable valid causal inferences to be made about the compared training methods. Examples of empirical evaluations of SIMNET[2] are the research of

1. Kraemer and Bessemer (1987).

2. Gound and Schwab (1988).

3. Brown, Pishel, and Southard (1988).

4. Test and Experimentation Command (TEXCOM) Combined Arms Test Center (1990).

Kraemer and Bessemer (1987)

Kraemer and Bessemer reported observations of U.S. platoons before and during the 1987 Canadian Armor Trophy (CAT) competition. They found a positive correlation (r = .53, n = 9) between the number of SIMNET battleruns completed by nine platoons and the scores obtained by those platoons in the CAT competition. With this small number of platoons, however, the correlation was statistically unreliable. A correlation coefficient of .53 must be based on a sample of at least 14 platoons to be judged significant at p = .05.

Kraemer and Bessemer also reported that the observed relation between platoons' number of SIMNET battleruns and CAT scores may have been entirely spurious. The spurious relation may have been caused by a single extreme platoon, which trained much more than others and happened also to obtain the highest CAT score. When the extreme platoon's score was removed from the analysis, the correlation between

---

[2]Our reviews of the four empirical evaluations of SIMNET are partly based, with permission of the senior author, on an unpublished manuscript by John E. Morrison, Eugene H. Drucker, and David Campshure. Their manuscript is entitled Devices and aids for training M1 tank gunnery in the Army National Guard (Alexandria, VA: Human Resources Research Organization, 1990). The reanalyses of data reported here were done by the present authors.

SIMNET training and CAT scores reduced to near zero. If the training effect on performance only emerged with a moderate or large amount of training, then deliberate manipulation of the amount of training was necessary to estimate the nature of the relation. Platoons that completed the needed intermediate numbers of battleruns were unavailable in the CAT sample. Noting their research was not specifically designed to estimate the contribution of SIMNET training to CAT scores, Kraemer and Bessemer concluded, "It is impossible to determine with certainty whether SIMNET training benefitted, reduced, or had no effect on the performance of the U.S. platoons in the CAT competition" (p. 28).

Gound and Schwab (1988)

Gound and Schwab pretested eight platoons on parts of three Situational Training Exercises (STXs), using MILES-equipped tanks at Fort Hood. Four of the eight platoons then trained with SIMNET for 50 to 52 hours each, and the other four platoons trained conventionally in the field at Fort Hood. After training, all platoons were tested on three similar STXs with the MILES-equipped tanks at Fort Hood. Observers scored performance as GO or NO-GO for 51 pretest tasks and 55 posttest tasks in the STXs, but tasks were not scored if they were not observed. Appropriate rank order tests were used to compare the percentage of tasks scored as GO on the pretest and at the posttest. The changes in this measure were not, however, compared between the two groups. Differences found between the pretest scores of the compared groups were declared to be statistically significant ($p = .057$); kinds of training (SIMNET and conventional) were thus confounded by between-group proficiency differences that existed before the experiment began. That confounding led Gound and Schwab candidly to assert that any differences between the compared groups' posttest scores could not be attributed to differences between SIMNET and conventional field training.

Brown, Pishel, and Southard (1988)

Brown et al. (1988) identified ten tasks that were common to SIMNET and to each of eight STXs. Using data from Gound and Schwab's platoons, Brown et al. (1988) reported a non-significant difference between the SIMNET-trained and field-trained groups' pretest scores on the ten tasks, but a significant difference between the compared groups' posttest scores. Brown et al. (1988) also analyzed data from the same tasks in a subsequent company team Army Training and

Evaluation Program (ARTEP) exercise performed two days after the STX posttest. They reported that the ARTEP scores of the SIMNET- and field-trained groups were not significantly different.

Brown et al.'s analyses used the Chi-square statistic inappropriately and should be reexamined. Each test was based on a 2 x 2 table, with the SIMNET- and field-trained groups as one dimension and the GO and NO-GO categories as the second dimension. The frequencies entered in the table were task counts combining all four platoons. Such data can be combined legitimately from different platoons for a Chi-square test only when they are homogeneous samples, that is, when the probability of a GO score is the same for all platoons and all tasks. Such an assumption could not be justified in Brown et al.'s research. The only generalization possible from the analysis as conducted concerns aggregated platoon data rather than individual platoon, and only one aggregation is available for analysis; that is, there are no replicates.

An appropriate procedure is to treat the tasks used by Brown et al. as test items and to treat the number of tasks rated GO on pretest, posttest, and ARTEP as scores for each platoon. The key questions for analysis are (1) whether the changes in scores (gains) from pretest to posttest differed between the groups with SIMNET or field training, and (2) whether the groups' scores differed on the tasks in the ARTEP. Results of a repeated-measures analysis of variance by the present authors showed that the observed group difference between these changes (group by test interaction) was not significant ($p = .205$). Although the average gain from SIMNET training could be larger than the gain from field training, the observed difference between gains may be attributed to normally expected performance variations from pretest to posttest. Analysis of the ARTEP scores also did not show a significant difference between the compared groups' scores ($p = .304$). Thus, even if the SIMNET training produced better scores on the posttest than field training did, there was no reliable evidence that the group difference was stable enough to transfer and cause a similar performance difference in the company-level ARTEP missions.

Furthermore, the small number of platoons in the SIMNET- and field-trained groups virtually precluded Brown, et al. (1988) from finding any significant differences. We estimated the chances of obtaining significant statistical results (power of the test) if the actual training effects were the same size as those observed by Brown et al. to be

less than 25% for both statistical tests reported here, based on the variability in scores obtained among the platoons.

TEXCOM Combined Arms Test Center (1990)

In TEXCOM's evaluation of SIMNET, nine armor platoons and nine mechanized-infantry platoons initially performed a movement-to-contact STX in the field. All platoons then received 2 days' orientation and 3 days' SIMNET training on STX exercises. The platoons then took a posttest, with task contents identical to the pretest's. Pretest and posttest exercises were conducted on different terrain, however, causing terrain effects on performance to be confounded with the pretest-posttest performance change. Scores on collective subtasks and individual tasks generally improved from pretest to posttest. The average gain[3] in percentage of subtask standards judged GO was statistically significant for both the armor and the mech-infantry platoons. Because they used no control groups, however, the TEXCOM researchers' results permitted no valid causal inferences, that is, inferences between the effect of SIMNET training and proficiency increases. All increases in scores may have been due to practice during the pretest or to easier posttest terrain, rather than to SIMNET training.

The TEXCOM research also assessed sustainment training in a novel way: GO scores on similar pretest and posttest items were taken as evidence that the intervening SIMNET training was sufficient to sustain task performance. Sufficiency is a weak criterion without evidence that some training is necessary. More rigorous tests of sustainment training would combine tests of necessity and sufficiency: How, for example, do the effects of sustainment training with SIMNET compare with the effects of no sustainment training? The more rigorous test of necessity seems especially germane as a

---

[3]The use of gain scores has a number of disadvantages: (1) Gain scores do not reflect where the compared groups began or ended; (2) groups that score lower on the pretest can achieve greater gains than groups that score higher; (3) gain scores compound the unreliability of the two scores from which they are derived; and (4) comparisons of gains for device- and conventionally-trained groups provide no indication of how much, if any, of the gains were ascribable to training (Boldovici, 1987). Harris (1963), Cronbach and Furby (1970), and Collins and Horn (1991) present acceptable alternatives to gain-score comparisons for estimating and analyzing change.

baseline for comparison inasmuch as training developers and researchers traditionally have regarded items passed on a pretest, not as springboards for demonstrating the sufficiency of sustainment training, but as indicators that no training on objectives pertaining to the passed items is required.

Additional Evaluation Data

In each of the cited research examples, various kinds of observations, instrumented data, and questionnaire responses were obtained to supplement the primary performance data. The supplementary data were analyzed by quantitative and qualitative methods to address secondary evaluation issues. The TEXCOM research, for example, gathered MILES-based casualty assessments and range of weapon engagement in the STX exercises. In relation to the SIMNET training, data were gathered on training time and simulator reliability, distance traveled, rounds expended, and soldiers' opinions about SIMNET realism, acceptability, and human factors.

The supplementary data sources provided useful information on the strengths and weaknesses of SIMNET and helped identify needed modifications. The supplementary data also provided insights into the value of exercise scenarios, uses of semi-automated forces, after-action review procedures, and the effectiveness of training methods. The cited research was, however, often constrained by narrowly conceived evaluation issues and limited resources. Too few rather than too many supplementary data consequently were collected, leaving many evaluation questions unanswered.

A common omission that is important for the interpretation of transfer effects in device evaluations is performance data from the device-based training exercises. In-device performance data obviously are needed to determine whether device training improves soldiers' performance in the device. If performance does not improve in the device, then little or no transfer of training can be expected; such a result simply provides evidence that the amount or kind of device training was inadequate to master the trained task. If performance in the device does improve during training but does not transfer, then what was learned in the device was somehow unusable to perform the task in the transfer situation. One possible interpretation is that the stimulus conditions presented in the simulation differed too much from those in the transfer situation, so the trained individual or unit failed to recognize when or how the responses that were learned should be performed to accomplish the transfer task.

Another interpretation is that the wrong responses were practiced and learned in the device, and a different set or form of responses was needed to perform the task in the transfer situation. Evaluation data should be comprehensive enough to provide a clear picture of what responses were practiced and learned in device training and what circumstances cued those responses, as well as what responses were performed under what conditions in the transfer situation.

Advantages of Empirical Evaluations

The chief advantage of empirical evaluations such as those described above seems to have been in using their results to address the four device-evaluation objectives mentioned earlier. Those objectives pertained to:

1. Support for inferences about the effect of CCTT training. Alluisi (1991), for example, cited Kraemer and Bessemer's results as an indication of the effect of SIMNET training on the U.S.' win of the 1987 Canadian Armor Trophy.

2. Budget justification. In a statement before the Senate Armed Services Committee on 21 May 1992, GEN (Ret.) P.F. Gorman seems to have referred to the TEXCOM (1990) study as demonstrating SIMNET's effectiveness as a sustainment trainer. (The transcript does not provide an exact citation; we inferred that the TEXCOM study was the cited one from Gorman's description of "a test of training transfer for nine armor platoons and nine mechanized infantry platoons" [p. 16].)

3. Compliance with acquisition, test, and evaluation regulations. The Gound and Schwab (1988) study mentioned above fulfilled a requirement for performing a Concept Evaluation Program test. The Brown, Pishel, and Southard (1988) study fulfilled a requirement for performing a Preliminary Training Device Study. And the TEXCOM (1990) study fulfilled a requirement for performing a Force Development Testing and Experimentation evaluation.

4. Recommending ways to increase simulator training capabilities. Each of the cited empirical evaluations identified some shortcomings in simulation and in training methods, and each of the evaluations recommended simulator modifications and ways to improve training. In many cases, similar findings were repeated in several reports, giving cumulative force to recommendations.

Disadvantages of Empirical Evaluations

The chief disadvantage of empirical evaluations is that limits on resources for examinations of transfer to field settings have caused compromises in research designs and execution. In the empirical evaluations of SIMNET mentioned above, the compromises were so severe that they precluded valid inferences about the effects of SIMNET training on soldiers' performance in the field. The compromises created evaluation flaws, which included:

1. Insufficient statistical power to demonstrate transfer differences that may have in fact existed between the compared groups.

2. Inadequate sampling, resulting in confounding training treatments with pre-experimental proficiency differences between groups or gaps in the values available for the sampled independent variables.

3. Inappropriate statistical analyses, sometimes leading to incorrect statistical inferences or to answers to the wrong evaluation questions.

4. Inadequate controls, which confounded the effects of uncontrolled variables with the training treatments and precluded estimates of the independent effect of SIMNET training on transfer scores.

5. Failure to collect data needed to properly interpret transfer results or needed to indicate ways to improve the simulation and its use for training.

The effect of compromises in evaluation designs, and especially the effect of insufficient statistical power, is to stack the evaluation deck in favor of finding no statistically significant differences between transfer scores of the compared groups.[4]

---

[4]For additional discussions of compromises in evaluation designs and their negating effect on valid inference, see Campbell (1957); Campbell and Stanley (1963); Cook and Campbell (1979); Horst, Tallmadge, and Wood (1975); Boldovici (1987); and Morrison (1990). Cohen (1988) presented a comprehensive compilation of power-analysis methods, and Morrison (1990) described procedures for determining the numbers of platoons necessary for demonstrating training effects.

Finding no statistically significant differences between the transfer scores of compared groups does not demonstrate that no differences exist. The probability of discovering between-group differences that do in fact exist (the probability of correctly rejecting the null hypothesis) is governed by statistical power, which increases with increased numbers of observations (scores), decreased variation among scores, and increased size of the actual effect. Morrison (1990) demonstrated that null results in many Army transfer experiments were caused by researchers' use of too few observations: Statistical power was so small as to make detection of substantial true differences between the scores of compared groups improbable. Only extremely large differences could be declared significant in the research Morrison reviewed. Implications for training-effectiveness criteria that require finding no differences between the transfer scores of groups using devices versus weapons in training are obvious: The easy way to bias evaluations in favor of finding no significant differences between the transfer scores of compared groups (SIMNET vs. weapons, for example, or CCTT vs. weapons) is to use small numbers of units in the comparison. The usual resource limitations that encourage experimentation with small samples thus make training-effectiveness criteria based on device-weapon equality easy to satisfy. Finding no differences between the scores of device- and weapons-trained groups leads to the erroneous conclusion that satisfies device advocates and seems to justify device acquisition: The tested device is "just as good as" the weapon for training.

The subtlety here is, of course, as implied earlier: Many findings of no difference between the scores of device- and weapons-trained groups are more likely to result from inadequate sample sizes than from the absence or small size of differences between the scores of compared groups. One of the tenets of modern statistics is that the nonexistence of a difference--that is, "equal effectiveness"--cannot be proved. Failure to reject the null hypothesis does not logically justify the conclusion commonly drawn from the result and does not support acquisition decisions. Clearly, device evaluation must be based on more appropriate criteria of training effectiveness than "equal effectiveness."

A reviewer reminded us that, "Calculation of confidence intervals is of course an alternative [solution] to the 'equal effectiveness' problem" (Dwight J. Goehring, personal communication, 21 September 1993). We agree and suggest that researchers who find no transfer differences between compared groups should routinely report confidence intervals and

explain in lay terms the relevance of confidence intervals to findings of no difference.

Confidence intervals for mean differences (Festinger & Katz, 1953) are an effective way of presenting statistical results, especially to support logical scientific conclusions that are subject to revision in light of additional evidence. Hypothesis tests, on the other hand, are more compatible with situations in which decisions about alternative courses of action must be made based on observed differences--decisions that, once made, are irrevocable. (A collection of articles discussing this issue is in Lieberman, 1971, pp. 115-174.) Confidence intervals and hypothesis tests are closely related. A confidence interval for a difference specifies the limits of a range of difference values that has a high probability (such as $p = .95$ for a 95% confidence interval) of including the true difference. If the hypothesis test on an observed difference between compared groups is nonsignificant, then the corresponding confidence interval will include the value of zero.

The main virtue of the confidence interval is that it displays the set of nonzero differences that are plausible given the data actually obtained. In contrast to simply showing an observed difference to be statistically significant or not significant, the confidence interval supports a conclusion such as "If the true difference is not zero, then it is unlikely to be less than X_1 or greater than X_2." If the interval is long, and the limits of the interval are distant from zero (i.e., many large differences are plausible), then any conclusion that the difference is truly zero (treatments "equally effective") or near zero becomes insupportable. Confidence intervals for many of the differences between alternative training treatments that device evaluators have declared "equally effective" (including device and weapons alternatives) would, we suspect, demonstrate those declarations to be in error.

In addition to reporting confidence intervals, evaluators who find no transfer differences between compared groups also should report the result of power analyses (Cohen, 1988). A power analysis done before an evaluation will estimate the probability that the planned experiment is capable of detecting group differences that are large enough to have important practical consequences. Such an analysis will require some way of estimating the variability of scores within groups. The estimate can be obtained from a prior experiment or a pilot study. The benefits of discovering inadequate power before the research begins are obvious: We

can change the design to obtain adequate power, or we can abort the evaluation to avoid wasting money.

If power analyses are not done before conducting an experiment, and no differences are found between alternative training treatments, then we suggest doing power analyses using data collected during the evaluation. Reporting the results of power analyses would help the research consumer decide whether the nonsignificant differences were likely to have resulted from small effects of treatments on the performance measures or from an evaluation design that was unlikely to detect major differences. Performing post-evaluation power analyses also would lead to accumulating lessons learned with implications for future evaluation designs.

Aside from their ability to fulfill test and evaluation regulations, comparisons of the effects of training with devices and with weapons are hard to defend. Even if we assume, erroneously, that failure to find differences in transfer due to training alternatives proves device- and weapons-training equally effective, what may we legitimately conclude? We may legitimately conclude that training with devices is as effective as training with an alternative whose effectiveness is unknown. No significant differences interpreted to mean equal effectiveness thus not only are grounded in statistical and logical misunderstandings, but also are uninformative. Declarations of equal effectiveness are, in addition, Pollyanna-like: As Sticha, Singer, Blacksten, Morrison, and Cross (1990) suggested, when no differences are found between the effects of training alternatives, we may as legitimately declare the alternatives equally ineffective as equally effective.

The reasons for comparing the effects of device- and weapons-training seem grounded in muddled thinking. Consider, for example, that the reason training devices are built is because we know in advance that training with devices is going to be cheaper than training with weapons systems. If the case for savings could not be made before a device is built, then the device would not get built. We also know in advance (1) that the parent weapons systems cannot be used for training on any but limited scales, and (2) that the limited scales will get more limited as the prices of weapons go up and the budgets for weapons go down. As if to belie those certainties, we conduct training-effectiveness evaluations by comparing the effects of training with devices to the effects of training with weapons--weapons whose infeasibility for widespread use in

training we knew in advance and was the reason for building the devices in the first place.

If saving money is our objective, would not economy dictate that we compare the effects of device training to the effects of training with less expensive alternatives rather than with more expensive weapons? And in cases of sustainment training--the CCTT's reason for being (TRADOC System Manager for CATT, 1991, p. 1-5)--why compare the effects of device-based training with any training at all? We should like to discover how device-based training compares as a sustainment medium, not to high-priced weapons alternatives whose infeasibility of use we know in advance, but to a no-cost alternative that could save money, namely, no training. The first two questions that should come to mind for evaluating sustainment training are (1) "Is the training better than nothing?" and (2) "If so, how much?" A long-term benefit of using no-training control groups is that eventually interested researchers and the U.S. Army would discover the conditions under which various amounts and kinds of sustainment training are and are not required.

The illogical nature of device-weapon comparisons is especially apparent when the test result shows the device to be less effective than the weapons system for training. This result does not mean that device training is worthless. To the contrary, the device's absolute training value can only be measured against a no-training control, to determine if the device training is better than nothing. The inference cannot therefore be that the device has no value for training, and the conclusion is irrelevant to the acquisition question. The important research issues are: (1) whether device training can save some portion of the more costly on-weapon field training, (2) whether device training can enable subsequent on-weapon training to start and finish at a higher level of performance, thereby increasing the return on investment from training with weapons, and (3) how the benefit of device training compares to the benefit of lower-cost training such as classroom instruction, map or terrain-board exercises, and computer simulations that are available and commonly used to prepare for weapons training or to substitute for weapons training that is unavailable or too costly.

Device acquisition decisions must be based on the relative costs of devices, weapons, other training media, and the trade-offs in effectiveness for various mixes of training media (Hoffman & Morrison, 1992). Current device-testing practices provide neither the kinds nor the amounts of data necessary for making such tradeoffs.

Demonstrating training effects using traditional device evaluation paradigms will become more difficult as we move from SIMNET evaluations with platoons to CCTT evaluations with company teams and battalions. That is because, all other things equal, the numbers of sampling units required for sufficient statistical power to demonstrate training effects remain the same whether the sampling unit is the individual, the crew, the platoon, or the company-team. We were unable to provide enough platoons to demonstrate training effects with SIMNET; how then shall we provide enough company teams (and combat-support complements) to demonstrate training effects with CCTT?

The consequences of continued adherence to traditional device evaluation methods will include additional spurious inferences about "equal effectiveness." Future evaluations must take a new tack to address the issues discussed here.

Another disadvantage of traditional empirical evaluations of the kind performed to meet acquisition regulations is that their one-shot character and the limits on evaluation resources guarantee the impossibility of controlling or randomizing the many variables that could affect evaluation outcomes. The one-shot character of device evaluations is in fact antithetical to the concept of research, that is, of searching again.

Lest our observations be dismissed as ivory-tower concerns, we emphasize that failure to control or randomize major variables that could affect evaluation outcomes yields evaluation outcomes that have no use in establishing training effectiveness (or ineffectiveness). There are no 80% solutions in device evaluations, because 80% solutions usually produce conclusions that are 100% invalid. We should rather have no evaluation outcomes than compromised evaluation outcomes. No outcomes leave uncertainties that can be guarded against by compensatory measures. Compromised outcomes mislead.

One-shot empirical evaluations are likely to yield misleading results for two reasons in addition to the impossibility of controlling or randomizing all input variables. The first additional reason is that the direction and slope of transfer functions are likely to change over time and intervening practice (Gagne & Crowley, 1948; Ellis, 1969); that is, soldiers' performance on early transfer trials is not a good predictor of performance on later transfer trials (Boldovici, 1980), and transfer can change over the time between device training and transfer testing. Such effects also can depend on the amount of training and

original learning with the device (Krueger, 1929). The measured amount of transfer may improve markedly on later transfer trials because of proficiency gained during earlier transfer trials involving aspects of a task that were missing from device training, or transfer may decline because the way of performing the task learned on the device is a partial solution that must be replaced by a new technique to reach higher performance levels. Positive transfer also can be reduced by forgetting. Negative transfer can even change to positive transfer, when aspects of performance that produced negative transfer shortly after device training are forgotten, leaving intact other aspects that contribute to positive transfer. One-shot studies typically do not manipulate the amount of device training, the amount of training in the transfer situation, or the intervening time. Results depend on the specific values of these variables, and the generality of all results is open to question.

The second additional reason that one-shot empirical evaluations are likely to mislead is suggested by Bessemer's (1990) findings with SIMNET: As instructors gained experience with SIMNET, they became more proficient in using SIMNET for teaching. In Bessemer's research, transfer from SIMNET training to field exercises began to emerge three months and five classes after the Armor Officer Basic Course classes first were given platoon-level exercises in SIMNET. Transfer continued to increase gradually in the subsequent five months and seven classes for which data were obtained. Similar effects are likely with other complex device systems, including the CCTT. Early transfer tests that pair new devices with inexperienced instructors are likely to yield lower transfer scores than would be obtained on later tests, when the instructors would have gained proficiency in using the CCTT for teaching. The effect of instructors' lack of experience on one-shot evaluation results is compounded by the fact that instructors may be given not much training before using the new device to train soldiers or units. And the focus of whatever instructor training is given will probably be on operating the device and its features from the instructor's station. Little guidance, much less training, is likely to focus on how instructors should train for effectiveness.

A final disadvantage of the kinds of empirical evaluations discussed above is that they rarely provide diagnostically useful information on secondary issues such as (1) performance data from device training to shed light on why soldiers' transfer scores were low or high or to examine tradeoffs with device alternatives and (2) supplementary information on visual perception, radio communications, task

practice conditions, or conduct of after-action reviews that can be used to modify the device or its use for improved soldier performance.

Analytic Evaluations

Analytic evaluations do not use scores that reflect trainees' performance in training or testing. Analytic evaluations also do not use statistical analyses to determine the effects of alternative training regimens on performance differences among groups. Unlike empirical evaluations, analytic evaluations are based on experts' analyses of similarities and differences between training devices and weapons systems, both in terms of the equipment and the battlefield operating environment. Analytic evaluations can be done using PC-based methods such as Rose and Martin's (1984) Device Effectiveness Forecasting Technique (DEFT), or Sticha, Singer, Blacksten, Morrison, and Cross's (1990) Optimization of Simulation Based Training (OSBATS), or any of various methods reviewed by Knerr, Nadler, and Dowell (1984). Paper-and-pencil checklist methods also can be used, as they were by Burnside (1990) and by Drucker and Campshure (1990) for analyzing the strengths and weaknesses of SIMNET. The outcomes of Burnside's and of Drucker and Campshure's analytic evaluations were descriptions of similarities and differences between the SIMNET modules and visual representations of the battlefield compared to the corresponding weapons systems operated in a field-training environment. Burnside and Drucker and Campshure also made educated guesses about the effects of the similarities and differences they observed on transfer to field exercises.

Burnside (1990)

Burnside developed and used a rule-based method to estimate which ARTEP Mission Training Plan (MTP) standards could be met and which subtasks and tasks could be performed in SIMNET. A standard was rated "highly supported" (H), for example, if it could be met entirely in SIMNET, with all actions realistically performed. A "highly supported" task was required to have a majority of subtasks rated H (including all critical subtasks), each subtask having a majority of H standards. The results of Burnside's analysis suggested that only 34% of battalion task force tasks, 29% of company team tasks, and 41% of platoon tasks were "highly or partially supported" by SIMNET, according to the rules defining those categories. Burnside derived clear recommendations for modifying SIMNET to improve coverage of

MTP task content for STXs. His recommendations also have direct implications for the design of CCTT.

Drucker and Campshure (1990)

Drucker and Campshure's work resulted in recommendations for improved ability to practice 25 drills, offensive missions, defensive missions, and special operations. Drucker and Campshure also presented strong inferences about how stimulus differences between SIMNET and tanks would affect the likelihood of transfer to soldiers' performance in the field. And uniquely among SIMNET researchers, Drucker and Campshure (1990) warned against the consequences of training with a device that has "partial capabilities":

> The majority of tactical activities . . . could not be performed as they would in the field. . . . Given the knowledge that the cues or the responses (or both) will be different from those occurring on the actual battlefield, the training developer must decide whether the training will be beneficial or not (pp. 54-55).

Advantages of Analytic Evaluations

The chief advantage of analytic evaluations is that their results can be used to meet the CCTT evaluation objective pertaining to recommendations for increasing the training capabilities of devices. Additional advantages of analytic evaluations are:

1. The results of analytic evaluations can be used to help meet the CCTT evaluation objective that pertains to justifying budgets, by identifying tasks and mission segments that are unlikely to be trainable with the device.[5]

2. Analytic evaluations are helpful in designing training strategies, because analytic evaluations identify

---

[5]Meliza (personal communication, February 1993) noted that analytic evaluations are better for identifying training that will not be supported or will be based on wrong stimulus-response relations than for identifying training that will be supported. If discriminative stimuli are missing from the device, for example, then the device is unlikely to support learning or rehearsal of responses appropriate to the missing discriminative stimuli.

what cannot be taught with the device. Analytic evaluations also identify tasks trainable on the device that cannot be trained in the field because of cost or safety considerations.

3. Analytic evaluations can be performed before prototype production, by using specifications and mock-ups.

4. The price of analytic evaluations is small compared to the price of empirical, transfer-of-training evaluations. Burnside's analysis, for example, cost less than $50,000, Drucker and Campshure's slightly more than $100,000. (The evaluation budget for CCTT is, according to a May 1993 GAO report, estimated at 15 to 19 million dollars.)

5. Analytic evaluations are useful in identifying issues and forming hypotheses for empirical investigation: What effect, for example, will the CCTT's inability to support certain tasks and mission segments have on sustaining the performance of previously qualified crews, platoons, and company-teams?

Disadvantages of Analytic Evaluations

Analytic evaluations have three disadvantages:

1. The information from analytic evaluations yields weaker inferences about the effects of device training than the inferences that would ensue from tightly controlled empirical evaluations with adequate statistical power. The results of analytic evaluations also cannot comply with regulations requiring the demonstration of transfer.

2. The persons who perform analytic evaluations must be familiar enough with weapons systems, unit standard operating procedures (SOPs), field operating conditions, and mission scenarios to be able to identify subtle differences between simulations and field practice. Those persons also must have a detailed understanding of device capabilities and operation. In addition, the persons performing analytic evaluations must have enough human-learning expertise to be able to make tenable inferences about the transfer effects of similarities and differences between devices and their parent weapons systems.

3. The results of analytic evaluations derive from analysts' expertise rather than from hard data. Analytic results may therefore be less credible in the eyes of device proponents and acquisition decision-makers than are empirical

results--even though no empirical evaluation of SIMNET has produced results sufficiently reliable to support valid inferences about the relation between SIMNET training and soldiers' performance in the field.

Lessons Learned from SIMNET Evaluations

Before indicating new directions for evaluating CCTT, we shall summarize the lessons learned from empirical and analytic evaluations of SIMNET.

Lessons from Empirical Evaluations

Lessons learned from empirical evaluations of SIMNET include:

1. Empirical evaluations of SIMNET incorporated compromises in research design that led to insufficient statistical power, inadequate controls, inappropriate analyses, and irrelevant comparisons. The flaws were so severe as to preclude valid inferences about the relation between SIMNET training and soldiers' performance in the field.

2. Compromises and flaws notwithstanding, the results of the empirical evaluations of SIMNET were used to support inferences about SIMNET's training effects on soldiers' field performance (Alluisi, 1991), to justify expenditures (Gorman & McMaster, 1992), and to meet test and evaluation regulations (Brown et al., 1988; Gound & Schwab, 1988; TEXCOM, 1990).

3. Inadequate statistical power in empirical evaluations of SIMNET was related to the use of too few platoons to detect training effects that may have in fact existed. To detect reliable training effects, empirical evaluations of CCTT will have to use more company-teams than the numbers of platoons used in evaluations of SIMNET.

4. The one-shot character of empirical evaluations of SIMNET precluded controlling or randomizing many extraneous variables that could affect evaluation outcomes.

5. The results of one-shot transfer experiments may mislead, because the direction and slope of transfer functions are subject to change with practice and over time after initial training, and because research with

inexperienced instructors using new devices is unlikely to reveal maximum transfer effects.

6. Empirical evaluations of SIMNET yielded little diagnostically useful information, that is, information about why units scored high or low, information about tradeoffs among training alternatives, and supplementary information that could be used to modify SIMNET or CCTT or its use to improve soldiers' performance in the field.

Lessons from Analytic Evaluations

Lessons learned from analytic evaluations of SIMNET include:

1. Analytic evaluations of SIMNET provided diagnostic information for changing device design and use that promises to increase training effectiveness.

2. Because they identify tasks and mission segments not supported by the device of interest, analytic evaluations are useful for meeting the CCTT evaluation objective that pertains to justifying expenditures and supporting budget requests.

3. The results of analytic evaluations are useful for designing training strategies. (See Morrison & Hoffman, 1988, for an example.)

4. Analytic evaluations are useful for forming hypotheses for empirical testing. What effect, for example, does sustainment training with a device that has only "partial capabilities" (Drucker & Campshure, 1990) have on qualified units' retention and transfer of previously learned tasks?

5. Analytic evaluations can be performed before prototype production, and they are less expensive than empirical, transfer-of-training evaluations.

6. Analytic evaluations provide information that yields weaker inferences about the relation between training and field performance than would tightly controlled empirical studies. (Tightly controlled empirical studies were, however, not feasible with SIMNET and promise to be less so with CCTT). The results of analytic evaluations also cannot comply with requirements for demonstrating transfer of training.

7. Conducting analytic evaluations requires persons with high levels of expertise in device operation, weapons employment, unit SOPs, field practices, mission scenarios, and human learning.

8. Because they rely on expertise rather than on experiments, analytic evaluations may be less credible to some than the results of empirical evaluations, even though no empirical evaluation of SIMNET produced results of sufficient reliability to support valid inferences about the relation between SIMNET training and soldiers' performance in the field.

Guideposts for CCTT Evaluation

The CCTT system will be the most complex training device ever to be developed, because it is a combination of several training devices operating together. Evaluating the training effectiveness of such a complex system poses a challenge that far exceeds the challenge in any previous device evaluation. As a DIS system, the CCTT must serve needs of many training customers simultaneously. With so many components and kinds of users, the CCTT cannot succeed in entirety or fail in entirety. Some parts of the system will work well and will satisfy the customers using those parts. Other parts will work less well and will produce various degrees of dissatisfaction. Because CCTT is a multi-component, interactive system, none of its parts can legitimately be evaluated as an independent component.

The approach to CCTT evaluation must be grounded in coherent general principles that reflect the magnitude of the evaluation problem and provide consistent intellectual guidance in the search for solutions. We suggest several CCTT evaluation principles here, not as the best or the last word on device evaluation, but as points to be considered by persons responsible for developing and updating device-evaluation policy.

The first principle is that the CCTT must be evaluated in systems terms. The value of the CCTT can only be judged in relation to its role in, and impact on, the total Army

training system and in turn in relation to the total Army.[6] As the core medium for simulated maneuver training in the Army's CATS, the CCTT must be viewed with other devices and means of training as parts of a continually evolving mix of training resources. The proper focus from this standpoint is on the CCTT's contribution to the mix: What part of the total CATS burden should the CCTT be asked to bear? Evaluations should address how the CCTT complements or supplements existing training alternatives to support and implement CATS while remaining within contemporary and future budgetary limitations. Evaluations also may indicate that the CATS existing at various times needs to be revised.

The second principle is that CCTT evaluation should be approached as a part of a larger program of Total Quality Management (TQM) applied to the Army training system and directed toward continuous improvement in training. Although an Army-wide TQM program for training does not yet exist, there will have to be one (Booher & Fender, 1990) or something much like it if the Army is to be capable of maintaining high readiness in the face of reduced training resources. Applying TQM to CATS and to CCTT and other CATS components throughout development and implementation will help move the Army training system in the direction of continuous improvement.

A reasonable assumption can be made from the outset that little about the first version of CCTT and how it is used initially will be perfect. The entire acquisition and implementation processes should be planned to encourage repeated upgrading and improvement. One of the advantages of the planned DIS modular architecture is that it lends itself to step-by-step, module-by-module improvement. Continuous feedback from evaluation results is the essential ingredient for defining requirements for modifications and for assessing the success of each modification.

The third principle follows from the second: The CCTT evaluation process must be planned as a continuous, institutionalized part of the TQM process. Collection and analysis of training input and outcome data must become an integral part of all Army training to enable continuous

---

[6]A reviewer commented that our focus on CCTT's relation to the total Army may be too narrow: "[Consider] evaluation also as applied to joint and combined operations . . . get other Services and some potential Allies involved. That's how CCTT will actually be used" (Frederick J. Brown, personal communication, 17 January 1994).

training quality improvement and quality assurance. Practices of continuous evaluation have become institutionalized at the CTCs; building in capabilities for evaluation to support the CCTT training process will help to move other parts of the Army training system in the direction of continuous, institutionalized evaluation.

The issues to be examined and comparisons to be made will change with CCTT's stages of development and stages of implementation. In the beginning, the emphasis should be on concept definition for aspects of the CCTT that radically depart from the SIMNET baseline system or that involve new capabilities. Early prototypes should be reconfigurable to allow exploration of alternative concepts for the "same" CCTT subsystem or component, and early user testing should examine the effectiveness of the variants. In addition, contractors should be encouraged to explore innovative solutions to CCTT problems or other system enhancements that are outside the mainstream of the CCTT program. As components firm up and become interoperable, higher levels of organization and more complex missions will become feasible, allowing more formal controlled research designs. Later, phased fielding schedules can be coordinated with schedules at the CTCs to take advantage of the one-time opportunity to compare samples of CCTT-trained units to other baseline units prepared for the CTC by the pre-CCTT training system. After fielding is complete, evaluation should focus on product improvements and on ways to improve CCTT's use for training all Basic Operating Systems at all echelons.

Recommended Methods for Evaluating CCTT

Various methods can be used to meet the CCTT evaluation objectives mentioned earlier and to overcome the deficiencies that attended empirical evaluations of SIMNET. Our recommended methods should not be regarded as exclusive alternatives. The recommended methods may be used at one time or another to address issues for which they are best suited, and they may be used in combination to provide converging evidence compensating for various weaknesses of each method used alone. The recommended evaluation methods are:

1. In-device learning experiments.

2. Quasi-transfer experiments.

3. Correlational research with archived data.

4. Efficient experimental designs.

5. Quasi-experimental designs.

6. Improved methods for documenting training.

7. Analytic evaluations.

In-device Learning Experiments

Important information can be gained by conducting experiments that assess learning with the device as a function of amounts of practice and as affected by training conditions. These experiments can be used to avoid or reduce costs of field evaluations. Within-device training and testing can be used for many special-purpose experiments to test hypotheses about training improvements. The results of in-device learning experiments are more likely than field experiments to yield useful diagnostic information, because of opportunities for tighter experimental control and for increased statistical power due to greater numbers of observations. Once efficient ways of training have been identified with in-device learning experiments, then transfer of training can be investigated without wasting resources on evaluating poor training methods.

Consideration also should be given to using CCTT exercises as gates as they are defined in CATS: Platoons and company-teams trained on the device should be able to demonstrate certain levels of proficiency with CCTT before proceeding to more difficult device exercises or to field exercises. The levels of proficiency on CCTT can be arbitrarily set at first and adjusted later in light of empirical verification. The underlying assumption here is that if soldiers cannot demonstrate proficiency in performing exercises in CCTT, then they are not ready to practice or to be tested on similar exercises in the field. That assumption is equivocal, inasmuch as soldiers' learning may improve in training but not transfer field settings. The assumption should therefore be tested. Field validation of gate tests should follow validation in quasi-transfer experiments as described in the next section.

Quasi-transfer Experiments

Quasi-transfer experiments can be used for low-cost evaluations that yield empirical and analytic results. Quasi-transfer experiments are ones in which training occurs

on the device of interest, and transfer is assessed using a different device or a reconfigured version of the same device. Examples are experiments by Witmer (1988) and by Turnage and Bliss (1990), who used quasi-transfer experiments in research with tank gunnery devices, and of Lintern (e.g., 1987), who used quasi-transfer experiments in research with fixed-wing aircraft training devices. Quasi-transfer experiments for the CCTT might include training given tasks and missions on CCTT and testing for transfer to new tasks and missions on CCTT.[7] Gate test validations might use the same tasks and missions as trained, but with different terrain and opposing-force (OPFOR) conditions.

The chief advantage of quasi-transfer experiments is in using them for research on how training and testing conditions affect retention and transfer of training, while avoiding costs incurred in field testing with weapons systems. Another advantage of quasi-transfer experiments is that they permit collecting repeated measures from repetitions of individuals' or units' performance in training and in testing. Test reliability, which is necessary for statistical power and valid causal inferences, increases with the number of items constituting the test and increases with the number of test scores averaged or otherwise combined to make a composite score. The scores from quasi-transfer experiments are therefore more likely to provide statistically significant results and grounds for valid causal inference than are the smaller numbers of scores typically available from field experiments. Quasi-transfer experiments are especially useful for examining issues that are central to the CCTT's training-effectiveness potential: To what extent, for example, can previously qualified platoons and company-teams "mentally fill in the blanks" while undergoing sustainment training on missions and tasks that the CCTT only partially supports?

---

[7]Wordsmithing may be all that is necessary to use the results of quasi-transfer experiments to comply with test and evaluation regulations. If not, then work should begin to change test and evaluation regulations on grounds that required evaluations are prohibitively expensive, cannot meet minimal standards for valid causal inference, serve only ticket-punching functions, and yield no useful diagnostic or training-effectiveness information.

Correlational Research with Archived Data

Correlational research of the kind performed by Hiller, McFann, and Lehowicz (1990) to examine OPTEMPO effects also should be considered. Correlational research can be performed using scores derived from the CTC data that are routinely collected and archived by the U.S. Army Research Institute (ARI) and scores from SIMNET or CCTT data that can be routinely collected and archived using ARI's Unit Performance Assessment System (UPAS). The UPAS provides a low cost, personal-computer system for collecting, displaying, and analyzing network data from SIMNET exercises (Meliza, Bessemer, Burnside, & Shlechter, 1992). An upgraded version of UPAS will be available for use with CCTT. SIMNET, CCTT, and other training data should be archived and used for continuing examinations of the relation between amounts, kinds, and costs of training inputs to the amounts and quality of transfer of units' training to CTCs.

Candidate test sites include Grafenwoehr, where company-sized organizations anticipated no Maneuver Rights Area (MRA) days in 1993 and where all company-sized infantry, armor, and cavalry units conduct SIMNET training twice per year (message from CINCUSAREUR Grafenwoehr/AEAGC-T, 21 December 1992). Additional test sites should be considered in light of criteria such as expected locations of early CCTTs, contiguity with CTC rotations, and anticipated differences in requirements and strategies for active and reserve components. The use of long-term correlational research is the only feasible way to relate soldiers' performance in CTCs to variations in CCTT training strategies, to changes in amounts and sequences of device and field practice, and to future evolutions of CCTT and DIS. Capitalizing on the training data available through the use of UPAS with CCTT can be facilitated by establishing a Center for DIS Lessons Learned and archives compatible with those now in use for the CTCs.

Designers of all evaluations aimed at relating CCTT training to field performance should consider the points made by Meliza (personal communication, October 92), namely, that field tests can be simplified by focusing on "training points" that cut across tasks. One method for deriving aggregate measures of common performance capabilities is illustrated by the work of Wheaton et al. (1980) on evaluating platoon batttleruns. Application of similar methods to field exercises done for CCTT evaluation would increase the number of times particular aspects of unit performance are measured, thus increasing statistical power by obtaining larger numbers of scores with no increase in the

number of test exercises. Examples of some training points Meliza mentioned are:

1. Upward communication: Vehicle and unit commanders' reporting critical information such as crossing phase lines, making initial contact, receiving indirect fire, and reporting friendly and OPFOR battle damage.

2. Downward communication: Leaders' disseminating information to subordinates, such as disseminating graphics, conducting briefbacks, and issuing FRAGOs.

3. Defensive maneuver: Units' adjusting positions in response to probable or actual OPFOR attack, changing firing positions and gun tube orientations, for example, moving to covered positions, and timely withdrawal to subsequent battle positions.

4. Offensive maneuver: Units' adjusting actions in response to terrain, using routes of advance that offer cover and concealment, for example, avoiding enemy fire sacks, assaulting from the last covered and concealed position before the objective, and maneuvering on the enemy's flanks.

Detailed listings and definitions of company-level performance measures of this kind are in Leibrecht et al. (1992).

Efficient Experimental Designs

Experimental designs other than the completely randomized design can be used to advantage in CCTT evaluations. The completely randomized design is relatively simple to set up and analyze and has been used frequently in training device tests (see research cited earlier). Completely random assignment of sampling elements (battalions, companies, platoons, crews or squads, individuals), however, often provides much less statistical power than other kinds of designs. These other designs involve special arrangements of treatment conditions and sampling elements that increase the stability of descriptive statistics (e.g., averages or proportions) and increase the power of tests based on such statistics. These arrangements control a portion of the variation between sampling elements, preventing that variation from contributing to the random variation among treatments.

Two basic kinds of designs are discussed in the sections that follow: (1) randomized block designs and (2) Latin

squares. A third method, the analysis of covariance (ANCOVA), also is discussed. ANCOVA is a correlational method, which may be used as an alternative to other designs or in conjunction with them. Many design types are essentially more complex variations of the basic ones, but they involve the same principles. Complex designs are described in texts applying to behavioral research (e.g., Kepple, 1991; Kirk, 1968; Meyers, 1979; Winer, 1971). Other design texts for research in industrial, agricultural, or biological applications should be used with caution, because their guidance may omit special considerations that arise in behavioral research with human groups or individuals.

The effects of violating the assumptions underlying the use of efficient experimental designs usually produce biased or uninterpretable results. We therefore recommend seeking expert statistical advice to plan arrangements of conditions and a valid analysis before data are collected.

Randomized Block Designs. Randomized block designs use groups (blocks) of sampling elements that are relatively similar to each other within each group but dissimilar between groups. In this context the groups are called blocks to distinguish them from treatment groups. Within each block, one or more sampling elements are randomly assigned to each treatment condition in the experiment. The effect of this arrangement is to prevent differences between blocks from contributing to the random variation in differences between treatments. A randomized block design can sometimes result in dramatic increases in statistical power. If, for example, half the variance among sampling units is associated with the blocking variable, this has almost the same effect on power as doubling the number of sampling elements in a completely randomized design.

The grouping that defines blocks in randomized block designs may be a result of a natural clustering or an association that results in similarity, such as genetically similar littermates sometimes used in biological experiments. Or the blocks may be deliberately arranged, by matching or stratifying subgroups of the sampling elements based on values or categories of a measured independent variable that is known or is assumed to affect the dependent variable, such as intelligence, which affects academic performance in educational research.

In the military, the echelons of command provide unavoidable natural groupings, such as companies within battalions or platoons within companies. The subordinate units within a higher-level unit often tend to be similar

both for internal and external reasons. Internal sources of similarity include shared unit traditions, command philosophy and climate, policies, and SOPs. External sources include common geographical location, facilities and resources, operational and training equipment, and training distractors.

Table 1 shows a randomized block arrangement with six battalions as blocks and one company within each battalion randomly assigned to one of four treatment conditions. In a training experiment that might be done with the CCTT, this kind of design would be used to analyze company-level measures of performance. Because each battalion is equally represented in each treatment by one company, any differences between battalions in average performance do not affect the differences between treatments.

The same kind of design arrangement also can be used if the experiment involves platoon-level performance measures. In that case, with three platoons per company, there are three measures of performance for each company that provide an independent estimate of the variation among platoons. Under certain conditions, the company variations can be combined with the platoon variations. In that case, the effective sample size becomes the number of platoons rather than the number of companies in the experiment.

Table 1

Example of a Randomized Block Design

	Treatment Conditions			
Battalion	1	2	3	4
1	Co C	Co B	Co D	Co A
2	Co B	Co C	Co A	Co D
3	Co B	Co A	Co C	Co D
4	Co A	Co D	Co C	Co B
5	Co C	Co A	Co D	Co B
6	Co A	Co B	Co D	Co C

The nature of command structures limits opportunities to use randomized block designs, because the number of treatments to be compared cannot exceed the number of elements within a block. With armor battalions as blocks, for example, four treatments use up the companies in a

battalion. Similarly, armor companies have three platoons, limiting designs with companies as blocks to three treatments. Similar but more complex kinds of designs--incomplete blocks--can be used for certain specific treatment numbers that exceed such limits.

The most common military example of deliberate grouping to form blocks is the use of the Armed Services Vocational Aptitude Battery (ASVAB) classification of mental ability categories. In this and other cases of deliberate blocking based on measured variables, the purpose may extend beyond increasing statistical power. The primary interest often is in testing the generality of the treatment effects; that is, are the effects the same for all blocks? If the effects are not the same, do the treatment effects vary systematically with the block variable? One might ask, for example, "Is a given simulator-based training program equally effective for squad leaders in all mental categories, or do the most effective conditions differ between the higher and lower categories?" Such questions can be answered using randomized block designs as well as with the ANCOVA described later.

When individuals constitute the sampling elements, having a large number of individuals available for selection makes it possible to match or group them to form blocks with little difficulty. With units as sampling elements, however, the small numbers that are usually available to support the research often will preclude effective blocking and prohibit the use of randomized block designs. Such designs will be useable for CCTT evaluations only if large numbers of units participate in the research.

Latin Square Designs. In their simplest form, Latin squares are similar to the randomized block design in one respect, because sampling elements again are grouped by some natural or measured blocking variable. The Latin square is, however, complicated by cross-classifying the same elements by a second natural grouping or measured blocking variable. Alternatively, the second variable can be some manipulated independent variable imposed on the sampling units. There may be little intrinsic interest in the effect of the second variable (e.g., a nuisance variable), but it must nevertheless be varied rather than held constant in the experiment. In an experiment with four battalions as blocks, for example, the second variable might be four observer teams used to evaluate performance of four companies doing exercises at the same time, where some constraint prevents the exercises from being done at different times. The cross-classification by two variables is done to simultaneously control the effects of both variables.

The cross-classification sets up a square matrix of cells with rows defined by levels of one variable and columns defined by levels of the second variable, as shown in Table 2. The sampling elements in a particular cell are those that combine one specific row-variable value with one specific column-variable value. The key restriction in forming a Latin square is that the number of levels for both the column and row variables must equal the number of treatments to be compared in the experiment. Treatments are then arranged in the cells so that each treatment appears only once in every column and once in every row.

The counterbalanced arrangement of the Latin square insures that each treatment condition is combined with all levels of the row variable and with all levels of the column variable. The design enables treatment effects to be estimated from the data unaffected by either row differences or column differences. Furthermore, variation associated with both variables is removed from the estimate of random variation, so that random differences between treatments are reduced, and the power of statistical tests is increased compared to the corresponding randomized block design.

Table 2

Example of a Latin Square Design

	Observer Team			
Battalion	1	2	3	4
1	B(Co C)	C(Co B)	D(Co D)	A(Co A)
2	C(Co B)	D(Co C)	A(Co A)	B(Co D)
3	A(Co B)	B(Co A)	C(Co C)	D(Co D)
4	D(Co A)	A(Co D)	B(Co C)	C(Co B)

Note. Letters before the company in each cell of the square refer to treatment conditions.

An important class of variations on the Latin square, one that is likely to be useful in CCTT evaluations, is the repeated-measure Latin square design. In this kind of Latin square, all the treatments are given in some sequence to each of the sampling elements, and each sequence forms one row in the square. The successive occasions at which the treatments are given form the column variable of the Latin square. An additional nuisance variable often will be confounded with

the occasion columns. Table 3 illustrates this kind of Latin square design. Four companies (from the same or different battalions) are assigned to the rows, one company with each sequence of treatments. Different scenarios for the same kind of mission exercise also are assigned to the columns, thus confounding scenarios with occasions.

Table 3

Example of a Repeated-measure Latin Square Design

	Scenarios and Occasions			
Sequence and Company	1	2	3	4
1--Company A	B	D	A	C
2--Company B	D	C	B	A
3--Company C	A	B	C	D
4--Company D	C	A	D	B

Note. Letters in each cell of the square refer to treatment conditions.

Repeated-measure Latin squares have several advantages over other kinds of designs for training and transfer experiments. First, repeated-measure Latin squares can be used with almost any number of treatments, because additional occasions can be added to form squares of any size. This avoids the limitations of other designs on the number of sampling elements that were imposed by the structure of units in echelons.

Second, repeated-measure designs tend to be powerful, because the variations among sampling elements produced by all the variables associated with particular elements are completely controlled. The random variations that remain to influence the treatment effects are often a small fraction of the variation in a comparable completely randomized or randomized block design.

Third, with relatively small numbers of sampling units the design can be expanded to include additional squares. If the treatment arrangements in these extra squares are selected correctly, a variable of interest can be used as the column variable instead of a nuisance variable. Such expanded multiple-square designs allow the effects of the column variable and its interaction with the treatment

variable inside the squares to be properly estimated and tested.

Repeated-measure designs with multiple squares are a convenient way to manipulate amount of training jointly with other treatments in evaluations of training effectiveness and transfer of training effects. With squares added to the one shown in Table 3, each with the scenarios arranged in certain specific orders in the columns, the expanded design would allow the examination of the effect of the amounts of training that accumulate as additional scenarios are completed. This is a kind of carryover effect that is a main point of interest in training experiments, rather than a problem for the statistical analysis. With the right design arrangement, the estimated effects would be uncontaminated by the differences in exercise or task conditions between scenarios. With performance measures from simulator exercises, increasing performance indicates within-device transfer; with a field test repeated after every simulator exercise, increasing field performance would show cumulative transfer of training.

A second way of manipulating amount of training can be used when the column variable represents different segments of a training program, such as different missions to be trained in a specific order. Because it would not be appropriate to rearrange the mission order in different squares, the amount of training would be used as the treatment variable. In order to vary the amount of training while preventing carry-over effects, each mission would be given equivalent amounts of training, while the position of a field test exercise within the mission training is varied. For example, all missions might be given six repetitions in simulator training. Units might be tested on each mission after 0, 1, 3, or 5 repetitions in sequences corresponding to the arrangement of Latin square letters A, B, C, D in their assigned rows. After completing each test exercise, the units would return to the simulator to complete the remaining number of repetitions for that mission before moving on to train the next mission in the sequence.

A discussion of additional considerations in the use of Latin squares is in Appendix B.

Analyses of covariance. The analysis of covariance (ANCOVA) combines correlational methods with other kinds of experimental designs. ANCOVA can be used when independent variables are present in the experimental situation that can be measured but that are difficult or impossible to control or to vary on purpose. The measured independent variable in

ANCOVA is called a covariate. One or several covariates may be used in the ANCOVA, but only the case with a single covariate and single dependent variable will be discussed here.

The ANCOVA method becomes useful when one or more covariates are suspected or known to have an effect on the dependent variable in the experiment. Usually a linear relation (regression) between a covariate and dependent variable is assumed in ANCOVA, but nonlinear relations also may be analyzed.

ANCOVA may be used for two different purposes. The first purpose applies when the researcher has little intrinsic interest in a covariate's effects but suspects the covariate may affect the dependent variable by increasing the amount of variation found among sampling units. This increase in sampling variability produces an undesirable reduction in statistical power. In such a case, ANCOVA is used to remove a portion of the variability in the dependent measure that is correlated with variation in the covariate. The reduced variation in the dependent measure makes statistical estimates more precise and increases the power of statistical tests for differences between treatment groups. The stronger the relation between the covariate and the dependent variable, the greater will be the increase in statistical power. In this case, the purpose of ANCOVA is to control and reduce statistically the effects of an extraneous variable that is not amenable to experimental control. ANCOVA also can be used with repeated-measures designs such as the one described in the previous section. With these designs, however, the covariate or covariates must be measured before each successive administration of the treatments.

This first kind of ANCOVA application is illustrated by an example with platoons randomly assigned to several simulator training conditions and given a posttest exercise in the field. We expect the experience level of the platoon leader to influence the platoons' posttest scores; our interest in leaders' experience derives, not from our intrinsic interest in leaders' experience as an object of research, but from our suspicion that the covariate may affect the dependent variable by increasing the amount of variation found among the sampled platoons' posttest scores. We decide, therefore, to measure months served as a platoon leader for use as a covariate. If months as a platoon leader proved to correlate strongly with platoons' posttest scores, then using ANCOVA would increase statistical power for testing the differences between groups.

The second purpose for ANCOVA occurs when the effects of the covariate are an object of interest in our research. Determining the effects of the covariate alone is of practical importance, and an increase in statistical power is a secondary benefit. We might want to establish the relation between simulator pretest and field posttest scores, for example, in order to predict posttest scores. A linear prediction equation then could be used to establish a policy of omitting simulator training when pretest scores are high enough to indicate that no additional training is necessary. Parallel equations in groups given different amounts of training could be used to predict how much training should be given to units with different levels of pretest performance.

Figure 1 illustrates this kind of situation with an example based on artificial data. Performance scores are linearly related to the pretest covariate in Groups A, B, and C, given 5, 3, and 1 days of simulator training, respectively. If the minimum required performance score is 60, then the linear relation in Group C shows that 1 day of training is insufficient regardless of the pretest score. The relation in Group B indicates that 3 days of training will be sufficient for those with pretest scores of 75 or greater. In Group A, more than 5 days will be required for units with scores of 25 or less. The differences in performance levels between groups show that each day of training adds about 10 points. In addition to this information, the ANCOVA statistical estimates and inferences are based on the much smaller variation among deviations of observed data points from the lines instead of the full variation of scores within groups.

Our interest most often will be in determining whether the effects of treatment conditions vary as a function of the value of the covariate. Systematic variation of effects on the dependent variable resulting from different combinations of treatments and covariate values amounts to a covariate-treatment interaction, meaning that the effects of the two variables are not independent of each other. This kind of effect usually shows up in ANCOVA as nonparallel linear relations between the covariate and the dependent variable in different treatments. Such an interaction is particularly important when the lines cross. Interactions of this kind are termed "disordinal" because the rank order of the treatments based on values of the dependent variable changes at different levels of the covariate. The line for Treatment A might be fairly flat, for example, crossing under a more steeply rising line for Treatment B at some intermediate value of the covariate. If the treatments are alternative training conditions, this sort of pattern would

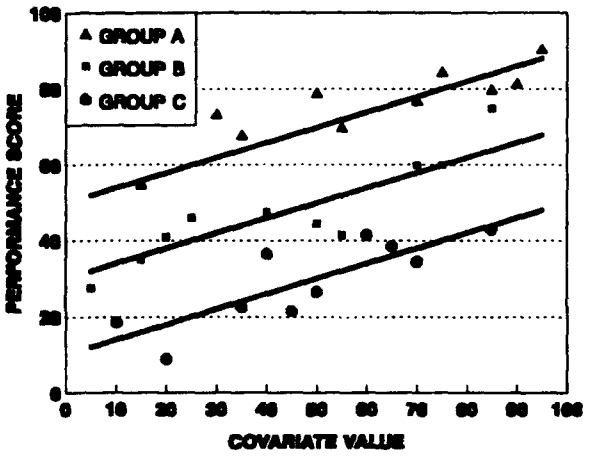


Fig. 1. Parallel linear relations within ANCOVA treatment groups.

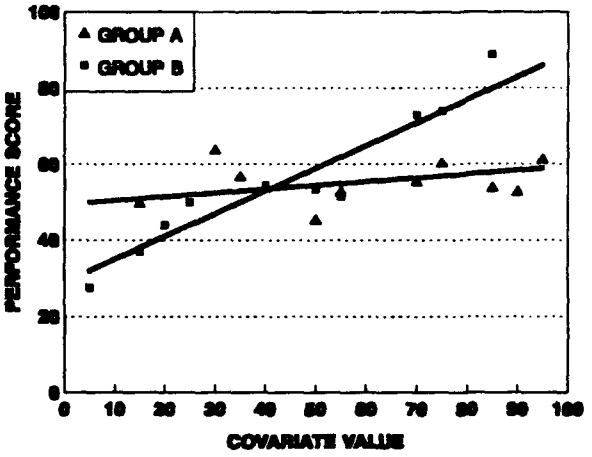


Fig. 2. A disordinal treatment group interaction with the ANCOVA covariate.

imply that Treatment A is better for units with low pretest scores, but that Treatment B is better for units with high scores. Figure 2 illustrates this sort of pattern, with the crossover occurring at a covariate value near 40.

One kind of covariate that should be useful in evaluations of the CCTT is pretest scores on tasks one echelon below the organizational level of the units to be trained. If the units sampled in the treatment conditions are battalions, then company-level tasks related to the same missions to be trained should be pretested, and if the units are companies, then platoon-level task should be pretested. The pretest scores then can be used in ANCOVAs performed both on CCTT training performance measures and on field performance measures. The objective of such analyses is twofold. The first objective is to determine if low levels of performance on the pretest tasks will prevent effective CCTT training and transfer below some score value. These results are important to define what proficiency must be attained at the lower echelon to make CCTT training worthwhile at the higher echelon. Such information forms a basis for effectively managing multiechelon training.

The second objective of using lower echelon pretest scores as a covariate is to determine whether there are pretest-treatment interactions that are disordinal. These

results are needed to determine how best to conduct training with units that come to the CCTT with widely varying capabilities to perform lower-echelon tasks. Such information can help maximize the benefits derived from a given investment in CCTT training.

In-device learning experiments and quasi-transfer experiments often use the same test to obtain pretest and posttest scores. Harris (1963) and Reichardt (1979) noted, however, that ANCOVA performed on the posttest scores as dependent variable with the pretest scores as the covariate ordinarily will be more powerful than the gain score analysis (Also see Boldovici, 1987, for a discussion of problems in using gain scores to estimate training effects.) The ANCOVA also will reveal treatment-pretest interaction effects: whether and how the treatment differences vary as a function of the pretest score. Advanced methods for the analysis of change are discussed in Collins and Horn (1991).

Three cautions must be observed in ANCOVA. They are discussed in Appendix C.

Quasi-experimental Designs

The validity of inferences is challenged by many possible sources of influence on research outcomes aside from the conditions controlled or manipulated by the experimenter. Cook and Campbell (1979) identified 26 categories of threats to valid inference, excluding those of a purely statistical nature. Many (but not all) of these threats can be controlled and excluded from consideration by the process of random assignment of units to the treatment conditions that are to be compared. In applied research, however, random assignment often is infeasible. Quasi-experimental designs are peculiar arrangements of treatments and units that offer protection from certain kinds of threats to inference without the requirement of randomization. When such designs are combined with specific prior knowledge or supplementary data bearing on the other possible threats, valid inferences may be reached even though full experimental control is lacking. The kinds of variables that must be considered are discussed by Cook and Campbell (1979).

Most of the quasi-experimental designs applicable to Army training and potentially useful for CCTT evaluations are ones that capitalize on the time-sequence for repetitive training events and on the cyclical turnover of groups in such events. For example, school courses may be repeated several times each year. Each class in that course has a new group of

soldiers who are given the same series of training events. Similarly, units are scheduled to appear at CTCs at regular intervals, and each unit may participate in the same or similar training events during its stay at the CTC site. In true experiments, groups with a randomly selected sample of classes or units would be assigned to each experimental treatment, and any unselected classes or units appearing during that same period of time would provide a control group. Cook and Campbell (1979) refer to this arrangement as an equivalent time-samples design, because the average conditions operative in the two sets of times are assumed to be roughly equal. A small two-group experiment of this kind is illustrated in Table 4, in which 12 classes or units appear at 12 consecutive times for training and testing. These sampling elements are randomly divided into experimental and control groups with six elements in each group. In most actual uses of this design a larger number of classes or units and times would be required for adequate statistical power.

Although the training or testing may vary nonrandomly across the series of points in time, for example, resulting from seasonal weather patterns or any other factors, randomization ensures (if the samples are sufficiently large) that the comparison between groups is not substantially biased by such differences. Calendar time might also be used as a covariate to reduce the influence of any systematic time trend in the dependent variable that results from conditions varying with time. Recording relevant conditions present at the various times would allow their possible effects to be examined directly.

Table 4

Example of an Equivalent Time-samples Design

	Time											
Group	1	2	3	4	5	6	7	8	9	10	11	12
Experimental	A		C	D		F			I		K	
Control		B			E		G	H		J		L

Note. Letters refer to consecutive classes or units.

The design can be strengthened by related prior-performance or pretest measures for each class or unit. For classes, such data will help control for class ability and other conditions that might uniquely influence the performance of that class. With units, a formal pretest is not necessary if measures from a prior appearance at the CTC are available for each of the units. These measures will help control for continuing conditions unique to each unit, such as home-station training resources or practices. Other supplementary data also are needed to examine what unit-unique variables have changed or remained constant between the first and second appearances at the CTC.

Consider also a similar design, but with systematic rather than random selection of units for the experimental and control groups. This forms a nonequivalent time-samples design, subject to all the hazards of interpretation common to other nonequivalent group designs. In CCTT evaluations with units, systematic nonrandom selection is almost inevitable, because the schedule of fielding to training sites will not be under control of the evaluators and is influenced by many practical considerations. Systematic selection risks confounding the treatment effect with the effects of other variables that differ between the groups. Without randomization, there is no basis for assuming that an experimental group composed of units using the CCTT for training is essentially the same as a group of units that do not yet have access to the CCTT. If a performance or transfer of training difference is found between the groups, the difference may have been caused by any one of a number of variables instead of the CCTT training. The burden of proof is on the researcher to rule out the effects of as many such other variables as possible. Prior training records and performance measures become critical to narrowing down the probable causes of differences between the transfer scores of compared groups.

One special case of the nonequivalent time samples design eliminates several threats to inference by varying the amount or degree of the treatment variable given to the units in the experimental group. This technique is analogous to the dose-response method used to establish the effectiveness of drugs or other treatments. With the CCTT, the evaluator might be given authority to manipulate the scheduling of units at the CTC, even though the fielding schedule cannot be controlled. Units then could be scheduled to appear after different time intervals from when the CCTT became available for training, and as the interval increased, the unit could complete a larger amount of CCTT training. Specific CCTT-trained units would be randomly assigned specific time intervals and

amounts of training. If CTC performance increased as a function of the interval and amount of training, then this effect probably could be attributed to the CCTT training. Only a few kinds of exotic threats to inference remain uncontrolled in this design. The randomization ensures that the time interval and amount of training is disassociated from other time-specific or unit-specific variables.

A different kind of nonequivalent time-samples design may be possible using training in school courses. If the CCTT is fielded to school sites on an all-or-nothing schedule, then in a particular course, no classes can train with the CCTT up to some point in time, and after the CCTT becomes available all classes train with the CCTT. This sequence forms an interrupted time series design as shown in Table 5. If training conditions are stable up to the point where the CCTT training begins, then performance measures plotted as a function of time should show a flat baseline with random variations around an average level up to the point of the change. After that point, any statistically reliable change in performance can be attributed to the effects of the CCTT if no other variables that might affect performance changed permanently at the same time.

Changes in performance might involve an abrupt jump to a new average level, a gradual increase to a new level, or a jump up with a later decline, as initial enthusiasm for the new device waned. Campbell and Stanley (1963, pp. 37-42) show examples and discuss interpretation of a variety of possible results in interrupted time series designs.

Additional data on a number of other kinds of variables are needed to rule out alternative explanations of an observed effect. Changes in the characteristics of

Table 5

Example of an Interrupted Time-series Design

	Time											
Group	1	2	3	4	5	6	7	8	9	10	11	12
Experimental							G	H	I	J	K	L
Control	A	B	C	D	E	F						

Note: Letters refer to consecutive classes or units.

individuals in the classes, for example, different dropout rates, changes in prior training, changes in the performance measures, or changes in test exercises are all possibilities to be guarded against.

An interrupted time-series design was used by Bessemer (1991) to examine transfer of training to field-exercise performance after the addition of two days of SIMNET training in the Armor Officer Basic Course. Instructor ratings of student performance in platoon leadership positions (with one to five student training-platoons in each class) established an average level of baseline performance based on 17 classes with 48 platoons over a period of 48 weeks. With SIMNET training before the field exercises, performance gradually increased in the subsequent 12 classes with 39 platoons over a 33-week period. This upward trend was interpreted as the effect of the instructors' gradually learning how to train more effectively using SIMNET. Although all other possible causes could not be eliminated with certainty, none of the other variables that were examined showed a similar gradual increase that would account for the performance effect.

The validity of inferences from the interrupted time series design can also be strengthened if field test exercises can be given after various amounts of CCTT training. This would involve interrupting the planned sequence of CCTT training at specific points for randomly selected classes, completing the test exercises, and then returning the class to the CCTT to complete their program of simulator training. The randomization prevents the amount of training from being systematically associated with any other changes that may occur after the introduction of CCTT training in the course. One advantage of this kind of experimental manipulation is that no class is deprived of any part of its scheduled training. The class schedule might, however, have to be extended to allow additional time for the test exercises.

The interrupted time series approach also can be used with units at the CTCs. To set up the interrupted time series, the time samples designs that were previously discussed must be combined with the collection of baseline data from additional control-group units appearing at the CTC before the time-sample part of the experiment begins. This would require that test exercises and performances measures to be used for CCTT evaluation be installed as a fixed portion of the CTC training and data collection system well in advance of the scheduled start of the evaluation research.

Improved Methods for Documenting Training

The lack of an effective Army-wide system for collecting and retaining training data is a major obstacle to research on the relation between training resources and practices and the resultant effectiveness of the Army training system. Although massive data bases on performance of units at the CTCs have been archived, finding data on what training was done and what proficiency levels were reached before the units arrived at the CTCs has proved to be difficult. The potential value of such data is illustrated by the research of Keesling, Ford, O'Mara, McFann, & Holz (1992). Although based on a small sample, their study demonstrated important possible effects of training resources available and training programs conducted at home station on performance of units at the National Training Center. More detailed data than were available to Keesling et al. (1992) will be needed to be able to examine the place of the CCTT in the CATS training mix. Detailed data include documentation on all training conducted at home station and field sites before and after units train at CCTT sites and at the CTCs. In addition, any records or data on performance in training exercises that can be obtained will be valuable. Units that are candidates for participation in CCTT evaluations should be contacted well in advance to install the necessary data collection and storage systems.

The establishment of a central electronic repository for all such training data and for test and evaluation data would facilitate broad participation by Defense and Army agencies with analytic capabilities and responsibilities, as well as by contracting and academic communities that have interests in investigating simulator training issues. Wide use of the data also would be assisted by providing Internet access to the archives. The agencies responsible for conducting CCTT evaluations will not have sufficient resources to address every issue that should be investigated.

Analytic Evaluations

In addition to empirical evaluations of the kinds outlined above, analytic evaluations of the kind performed by Burnside (1990) and by Drucker and Campshure (1990) for SIMNET should be continued. Analytic evaluations will continue to yield strong inferences about the extent to which various missions and tasks will be trainable with CCTT and to what levels. Those kinds of inferences are prerequisite to specifying the sequences and mixes of CCTT training and field training that we hope will form the core of training

strategies defining the role of CCTT in accomplishing CATS objectives.

The results of Burnside's and of Drucker and Campshure's SIMNET analyses also can be used as-is: By comparing the instructional strengths and weaknesses identified for SIMNET to the design specifications for CCTT (including QuickStart), the analytic evaluations will yield recommendations for product improvements based on value added, in terms of task coverage, for each. The resulting priorities for CCTT improvements can be weighed against the costs of making each recommended change.

Other recently developed methods, for example, for concept analysis of simulator modifications (Plott, LaVine, Smart, & Williams, 1992) and for training tradeoff analyses (Hoffman & Morrison, 1992) hold promise as useful additions to the methods used earlier.

Conclusions and Recommendations

Reviewing the lessons learned from SIMNET evaluations led us to conclude that:

1. One-shot empirical evaluations of the kind typically used to meet regulations should be avoided because of high cost and inability to meet any CCTT evaluation objectives, including the causal-inference objective aimed at supporting device acquisition decisions. Results of such evaluations provide an unreliable and usually misleading basis for acquisition decisions.

2. Analytic evaluations should be continued and expanded because they produce low-cost information that can be applied to improving device design and use. Analytic evaluations also produce information that can be used in budget justifications and that help to design training strategies. Analytic evaluations identify issues and hypotheses for empirical testing and are less expensive and more informative than one-shot transfer-of-training evaluations.

3. Empirical evaluations other than the one-shot variety performed in response to regulations for SIMNET should be sought to support the two CCTT evaluation objectives that pertain to making valid inferences and to complying with acquisition, test, and evaluation regulations.

4. Empirical evaluations of CCTT should be designed to correct flaws that characterized empirical evaluations of

SIMNET, namely, insufficient statistical power, inadequate controls, inappropriate analyses, and comparisons directed at irrelevant issues.

5. Evaluation methods appropriate for use with CCTT include in-device learning experiments, quasi-transfer experiments, correlating scores achieved in SIMNET or CCTT training with scores obtained during rotations at Combat Training Centers, efficient experimental designs (randomized block designs, repeated-measures Latin squares, and analyses of covariance), quasi-experimental designs, improved methods for documenting training, and analytic evaluations.

Our recommendations are

1. Evaluations should address how the CCTT complements or supplements existing training alternatives to support and implement CATS, while remaining within contemporary and future budgetary limitations.

2. CCTT evaluation should be a part of a larger program of Total Quality Management (TQM) applied to the Army training system and directed toward continuous improvement in training.

3. The CCTT evaluation process should be incorporated as a continuous part of the TQM process.

REFERENCES

Alluisi, E. A. (1991). The development of technology for collective training: SIMNET, a case history. Human Factors, 33(3):343-362.

Bessemer, D. W. (1991, January). Transfer of SIMNET training in the Army Officer Basic Course (ARI Technical Report 920). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A233 198)

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. (1980). Some problems in evaluating training devices and simulators. In Proceedings: 22nd Annual Conference of The Military Testing Association (pp. 239-260). Toronto, Ontario, Canada.

Booher, H. R., & Fender, K. (1990). Total quality management and MANPRINT. In H. R. Booher (Ed.), MANPRINT: An approach to systems integration (pp. 21-53). New York: Van Nostrand Reinhold.

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.

Burnside, B. L. (1990, June). Assessing the capabilities of training simulations: A method and Simulation Networking (SIMNET) application (ARI Research Report 1565). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A226 354)

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. Psychological Bulletin, 54, 297-312.

Campbell, D. T., & Stanley, J. C. (1966). Experimental and quasi-experimental designs for research. Chicago: Rand McNally.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Collins, L. M., & Horn, J. L. (Eds.) (1991). Best methods for the analysis of change. Washington, DC: American Psychological Association.

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design and analysis issues for field settings. Boston: Houghton Mifflin.

Cronbach, L. J., & Furby, L. (1970). How should we measure "change"--or should we? Psychological Bulletin, 105, 317-327.

Department of the Army (1986). Systems acquisition policy and procedures (AR 70-1). Washington, DC.

Drucker, E. H., & Campshure, D. A. (1990, June). An analysis of tank platoon operations and their simulation on simulation networking (SIMNET) (ARI Research Product 90-22). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A017 009)

Ellis, H. C. (1969). Transfer and retention. In M. H. Marx, Learning: Processes Vol. 1. New York: MacMillan.

Festinger, L., & Katz, D. (1953). Research methods in the behavioral sciences. New York: Drysden Press.

Gagne, R. M., & Crowley, M. E. (1948). The measurement of transfer of training. Psychological Bulletin, 45, 97-130.

Gorman, P. F., & McMaster, H. R. (1992, May). The future of the armed services: Training for the 21st century. Statement before the Senate Armed Services Committee. Washington, DC.

Gound, D., & Schwab, J. (1988, March). Concept evaluation program of Simulation Networking (SIMNET) (TRADOC TRMS No. 86-CEP-0345, ACN 86299). Fort Knox, KY: U.S. Army Armor and Engineer Board.

Harris, C. W. (Ed.) (1963). Problems in measuring change. Madison, WI: University of Wisconsin Press.

Hiller, J. H., McFann, H. H., & Lehowicz, L. G. (1990). Does OPTEMPO increase unit readiness? An objective answer. In Army Science Conference Proceedings, II (pp. 171-180).

Hoffman, R. G., & Morrison, J. E. (1992, April). Methods for determining resource and proficiency tradeoffs among alternative tank gunnery training methods (ARI Research Product 92-03). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A250 867)

Horst, D. P., Talmadge, G. K., & Wood, C. T. (1975). A practical guide to measuring project impact on student achievement (Contract No. DEC-0-73-6662). Washington, DC: U.S. Department of Health, Education, and Welfare.

Keesling, J. W., Ford, P., O'Mara, F., McFann, H., & Holz, R. (1992, June). The determinants of effective performance of combat units at the National Training Center (Final Report, Contract No. MDA903-86-R-0705). Presidio of Monterey, CA: PRC, Inc. and HumRRO, Inc. (Available from Chief, ARI Field Unit-Monterey, P.O. Box 5787, Bldg 104, Presidio of Monterey, CA 93944-5011.)

Kepple, G. (1991). Design and analysis: A researcher's handbook (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Kirk, R. E. (1968). Experimental design: Procedures for the behavioral sciences. Belmont, CA: Wadsworth Publishing Company, Inc.

Knerr, C. M., Nadler, L., & Dowell, S. (1984). Training transfer and effectiveness models. Alexandria, VA: Human Resources Research Organization.

Kraemer, R. E., & Bessemer, D. W. (1987, October). U.S. tank platoon training for the 1987 Canadian Army Trophy (CAT) competition using a Simulation Networking (SIMNET) system (ARI Research Report 1457). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A191 076)

Krueger, W. C. F. (1929). The effect of overlearning on retention. Journal of Experimental Psychology, 12, 71-78.

Leibrecht, B. C., Kerins, J. W., Ainslie, F. M., Sawyer, A. R., Childs, J. M., & Doherty, W. J. (1992, April). Combat vehicle command and control systems: I. Simulation-based company-level evaluation (ARI Technical Report 950). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A251 917)

Lieberman, B. (Ed.) (1971). Contemporary problems in statistics. New York: Oxford University Press.

Lindquist, E. F. (1953). Design and analysis of experiments in psychology and education. Boston: Houghton Mifflin.

Lintern, G. (1987, May). Perceptual learning in flight training. Paper presented at the Basic Research In-Process Review, U.S. Army Research Institute for the Behavioral and Social Sciences, Princeton, NJ.

Meliza, L. L., Bessemer, D. W., Burnside, B. L., & Shlechter, T. M. (1992, July). Platoon-level after action review aids in the SIMNET unit performance assessment system (UPAS) (ARI Technical Report 956). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A254 909)

Meyers, J. L. (1979). Fundamentals of experimental design (3rd ed.). Boston, MA: Allyn and Bacon.

Morrison, J. E. (1990, January). 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)

Morrison, J. E., & Hoffman, R. G. (1988, March). Requirements for a device-based training and testing program for M1: Volume 2. Detailed analysis and results (ARI Research Product 88-03). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A196 365)

Plott, C. C., LaVine, N. D., Smart, D. L., & Williams, G. S. (1992, April). Concept analysis for simulation modifications methodology (ARI Research Report 1613). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A250 947)

Reichardt, C. S. (1979). The statistical analysis of data from nonequivalent group designs. In Cook, T. D. & Campbell, D. T. (Eds). Quasi-experimentation: Design and analysis issues for field settings (pp. 147-205). Boston: Houghton Mifflin.

Rose, A. M., & Martin, A. M. (1985, June). Forecasting device effectiveness: III. Analytic assessment of DEFT (ARI Technical Report 681). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A160 029)

Sticha, P. J., Singer, M. J., Blacksten, H. R., Morrison, J. E., & Cross, K. D. (1990, September). Research and methods for simulation design: State of the art (ARI Technical Report 904). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A230 076)

TEXCOM Combined Arms Test Center. (1990, August). Close Combat Tactical Trainer (CCTT) force development testing and experimentation. (TCATC Test Report No. FD 0200, RCS ATTE-3). Fort Hood, TX.

TRADOC System Manager for CATT (1991, November). Combined Arms Tactical Trainer Master Plan (draft). Fort Knox, KY.

Turnage, J., & Boss, J. P. (1990, October). An analysis of skill transfer for tank gunnery performance using TOPGUN, VIGS, and ICOFT trainers (ARI Technical Report 916). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A231 156)

U.S. Army Training and Doctrine Command (1985). The TRADOC training effectiveness analysis (TEA) system (TRADOC Regulation 350-4). Fort Monroe, VA.

U.S. Army Training and Doctrine Command (1987). User test and evaluation (TRADOC Regulation 71-9). Fort Monroe, VA.

U.S. General Accounting Office (1993, May). Simulation training (GAO/NSIAD-93-122). Washington, DC.

Wheaton, G. R., Allen, T. W., Johnson, E., Boycan, G. G., Drucker, E. H., Ford, J. P., & Campbell, R. C. (1980, May). Methods of evaluating platoon battlerun performance (ARI Technical Report 78-A-24). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A061 369)

Winer, B. J. (1971). Statistical principles in experimental design (2nd ed.). New York: McGraw-Hill.

Witmer, B. G. (1988, May). Device-based gunnery training and transfer between the VIGS and the UCOFT (ARI Technical Report 794). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A197 769)

Appendix A

Need for Statistical Analysis

Measures of performance typically vary among sampling elements (individuals, teams, or units), and often the variations are large. When sampling elements are selected and randomly assigned to groups given different training experiences, the variations among elements cause average scores to vary among groups quite apart from any difference produced by the compared training conditions.

Figure A-1 shows two theoretical normal distributions of scores. Such distributions are often assumed to provide a model of the relative frequencies of all possible scores that could be selected under specific conditions. For example, the left-hand distribution might model the possible scores for sampling elements in a control condition with an average value (mean, represented by the symbol μ) of 100. The scores are most frequent around 100 and become relatively infrequent below 70 and above 130. The right-hand distribution has the same shape but with a mean of 110 and with most scores between 80 and 140. This might represent the possible scores for the same sampling elements if they are measured in an experimental condition. The main point is that the scores vary widely over a substantial range in both the control and the experimental conditions, with considerable overlap between the two distributions of scores. The standard deviation, which is a statistic, σ, that measures the degree of spread among scores, is 10 for each of the two distributions.

When groups of scores are randomly sampled from the same distribution, each group has a unique combination of scores and a different mean. The differences between means are the result of the random selection process that produces different sets of scores in each group. The variability among means therefore is sometimes said to result from chance rather than being caused by some effect of differing conditions among the groups. Figure A-2 shows the actual frequencies of groups of eight scores obtained at random from the distribution shown in Figure A-1 with a mean equal to 100. The scores vary around 100 in Groups 1, 2, and 3, and the mean scores for those groups are near 100. In Group 4, the scores fall mostly below 100, and the mean for this group is closer to 90 than to 100. It would be easy to think that the scores came from a distribution with a mean near 90 instead of from the actual distribution with a mean of 100.

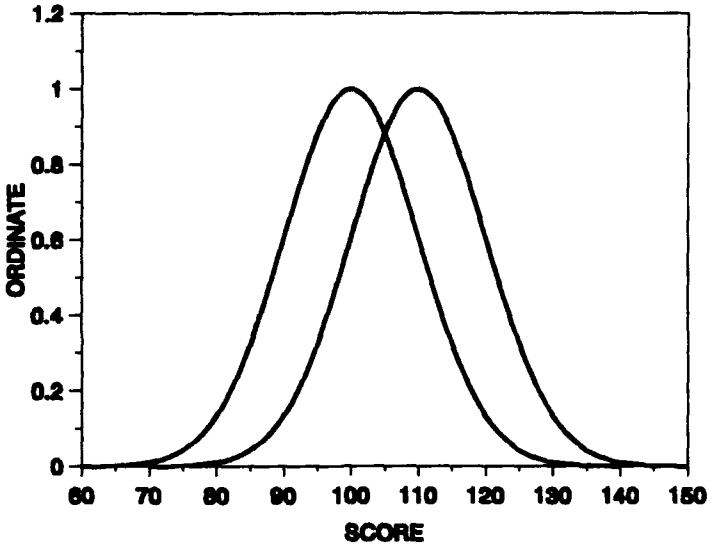


Figure A-1. Normal score distributions with means at 100 and 110.

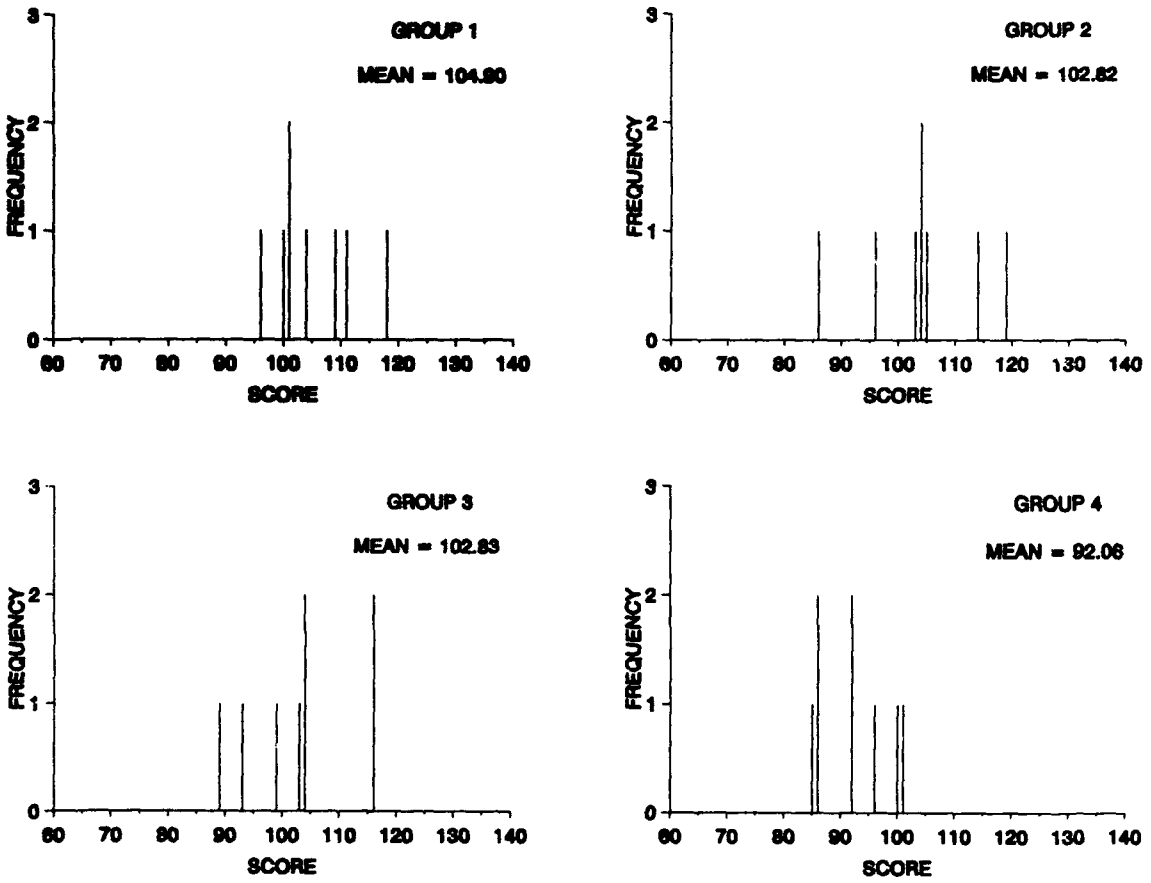


Figure A-2. Random samples of eight scores in four groups.

The distributions of scores resulting from research are estimates of the actual distributions of scores; that is, the actual distributions are not known. The means of the score distributions for control and experimental conditions might be exactly the same, if the experimental condition had no effect at all. Or the mean for the experimental condition might be larger or smaller than the mean for the control condition by any amount. When research uncovers an actual difference between means of two groups, a problem of interpretation immediately confronts the researcher. Could the observed difference be a result of chance alone, or is it more likely that the difference reflects a real difference between the distributions from which the scores were sampled? Statistical analyses use evidence obtained from the samples of scores to help answer such questions, enabling the researcher to draw appropriate inferences about the unknown distributions.

One form of statistical analysis uses a null hypothesis test. Such a test estimates the probability that a difference between an experimental group's and a control group's mean scores as large or larger than the difference actually obtained could have resulted solely from natural variations among the elements sampled for each group. The probability estimate is computed assuming that the real difference between the compared groups' scores is zero (the null hypothesis, $\mu_1 - \mu_2 = 0$) except for the difference caused by sampling variability. The alternative hypothesis is that the true difference is not zero, that is, $\mu_1 - \mu_2 \neq 0$.

If the estimated probability of the observed difference between the groups' mean scores is small, then the observed difference is attributed to effects caused by the training conditions instead of sampling variation. In this case, the null hypothesis (no difference between the compared group population means because the experimental treatment hypothetically does not affect training in any way different from the treatment of the control group) is rejected, and the observed difference between the compared groups' scores is said to be "statistically significant." The criterion for rejecting the null hypothesis is set in advance at a small value, usually $p = .05$ or $p = .01$, which is called the "level of significance" and is symbolized by α. This α-value is the level of risk accepted by the researcher of erroneously rejecting a null hypothesis that happens to be true. This kind of incorrect outcome of the hypothesis test is termed a "Type I" error.

On the other hand, if the probability estimated by the hypothesis test is larger than the level of significance, then the observed difference is attributed to the effects of sampling variation ("chance") instead of the training conditions. In this case, the null hypothesis is not rejected, and the observed difference between the compared groups' scores is said to be statistically nonsignificant. This conclusion might be in error if the alternative hypothesis is correct; that is, the real difference between the actual distribution means is not zero. This kind of incorrect outcome is termed a "Type II" error, and the probability of this error is represented by β. Statistical hypothesis tests are designed to minimize β, given the preset value of α. The actual value of β also depends on the number of scores in each group and on the sampling variability (σ) in each group. The value of β varies inversely with the size of the difference between means, $\mu_1 - \mu_2$, and has a different value for every possible difference.

The probability of correctly detecting an actual difference between distribution means, that is, of not making a Type II error, is $1-\beta$ and is called the "power" of the test. The test power is small for small differences, with $1-\beta$ approaching the value of α near zero difference. The power increases as the difference increases, approaching a probability of 1 for very large differences.

It is important to note that accepting the null hypothesis does not prove that null hypothesis is true, any more than rejecting the hypothesis proves that the null hypothesis is false. Accepting the hypothesis simply means that the observed difference was not large enough to provide an acceptable risk of making a Type I error. In practical terms, this result implies that we will act as if the null hypothesis is true until better evidence comes along, knowing that the chances of a Type II error have been minimized as far as possible under the conditions of the experiment. Rejecting the hypothesis simply means that the observed difference was large enough to make the risk of a Type I error acceptable. In this case, we can act as if there is a true difference with only a small chance of being wrong. However, we should also keep in mind that the hypothesis test does not establish exactly how big the actual difference might be, because the observed difference also will vary around the real difference in repeated experiments. Determining what possible differences are plausible in light of the observed differences requires using additional statistical estimation procedures such as confidence intervals.

Appendix B

Latin Squares: Additional Discussion

The negative side of Latin square designs is that special conditions and assumptions must be met for statistically valid estimates of effects and tests of significance. Although Latin square designs balance the overall (main) effects of the row, column, and treatment variables, the two-way interactions (row by column, row by treatment, and column by treatment) may be confounded with the estimates of effects and of random variation. The treatment effects will, for example, include variation from the row by column interaction effects. Such contamination may or may not be a problem for the statistical analysis and conclusions depending on specific assumptions that can be made about the nature of the variables and their interactions.

The possible problems with a Latin square design are eased if the design is replicated with additional sampling elements in each cell. With the design in Table 5, two battalions could be assigned to each row of the square, placing two companies in each cell. Such replication allows statistical tests to be performed to detect the existence of two-way interaction effects, and in some cases proper estimates of random variation can be formed even when the interactions are present. If the design is used with platoon-level measures, the platoons within each company provide the necessary replication without any additional companies.

An additional complication in Latin squares designs is that multiple measures taken from the same sampling elements are not independent. Different methods of statistical analysis have to be used that depend on the pattern of intercorrelations among the columns. In addition, the effect of a treatment itself may be affected by the treatments that precede it. Such effects are termed "carry-over" effects from the prior treatments, and they can bias estimates of treatment effects and invalidate the conclusions to be drawn from the results. Special experimental and statistical techniques may be required to prevent or correct for carry-over effects.

Appendix C

ANCOVA Cautions

Three cautions must be observed when using ANCOVA. First, the independent variables manipulated in the experiment cannot be allowed to influence the values of the covariate. This condition is usually ensured by measuring the covariate before the treatments are administered.

The second caution is that the covariate cannot be allowed to affect the nature of the independent variable that is applied to the sampling units or the test used to measure performance. Such effects may be subtle and difficult to prevent. For example, a training treatment that improved platoon performance could make the training exercises seem much easier, especially for the best platoons with the most experienced leaders. An exercise controller might inadvertently vary the exercise conditions to make them more difficult for the best platoons in that treatment group, while not doing so in other treatment groups. Similarly, a controller might make the test exercise conditions more difficult for the best platoons. If the posttest scores were reduced for those platoons by an increase in training or test difficulty, the posttest difference between treatment groups would be smaller as platoon leader experience increased, leading to an incorrect conclusion. If such side-effects of the covariate are difficult to prevent, they should at least be made detectible by documenting or measuring aspects of the treatment and test conditions that were administered whenever variations are possible.

Third, the sampling units should be randomly assigned to treatment groups. When intact groups of sampling units (e.g., companies or battalions when platoons are measured) are used to form treatment groups, any effects produced by preexisting differences between the intact groups are entirely confounded with the effects of the independent variable. ANCOVA cannot fully correct for this confounding, although it has been used for that purpose (see Kirk, 1968, pp. 455-458, for an example). One reason is that the covariate by itself does not completely measure the group differences that may affect the dependent variable. An ANCOVA can therefore remove only a portion of the confounded effect. With rare exceptions, valid inference requires random assignment to ANCOVA groups. This fact has been known for a long time (see Lindquist, 1953, pp. 328-330, for example), but many statistical and design texts do not

emphasize the point. The problem of drawing valid inferences from ANCOVA designs that use intact groups or nonequivalent control groups is treated in full by Cook and Campbell (1979). Such confounded arrangements should be avoided except when no other alternatives are possible.