The Elements of Training Evaluation

BY

JOHN A. BOLDOVICI

DAVID W. BESSEMER

AMY E. BOLTON

DISTRIBUTION STATEMENT A
Approved for Public Release
Distribution Unlimited

The U.S. Army Research Institute
for the Behavioral and Social Sciences

20020619 016

REPORT DOCUMENTATION PAGE

Form Approved
OMB No 0704 0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	April 2001	Book (Final) Jan 99 - Jan 01

4. TITLE AND SUBTITLE

The Elements of Training Evaluation

5. FUNDING NUMBERS

65803 D730
62785 A790

6. AUTHOR(S)

John A. Boldovici (U.S. Army Research Institute), David W. Bessemer (U.S. Army Research Institute), Amy E. Bolton (Naval Air Warfare Center - Training Systems Division)

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

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

8. PERFORMING ORGANIZATION REPORT NUMBER

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

U.S. Army Training and Doctrine Command
Fort Monroe, VA

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)*

This book addresses characteristics of field trials and characteristics of ratings necessary for making valid inferences about training effects, training capabilities, and proficiency. Chapter I describes and rebuts common rationalizations for conducting training evaluations that permit no valid inferences about training effects and for evaluation-reporting practices that preclude estimating the extent to which evaluation findings permit valid inferences about training effects. Chapter II presents elementary rules of evaluation design and analysis. These rules apply for the most part to the design of field trials and to the analysis and interpretation of data from field trials. Chapter III deals with advantages and disadvantages of ratings and with rules for their use. The kinds of ratings addressed are those used in the U.S. military for estimating the training capabilities of new training and for individual and collective performance appraisal. The rating rules describe ways to elicit reliable and therefore potentially valid ratings from which valid inferences may be made about the effects of training. In Chapter IV we suggest that, in light of the consistent failure of Army training evaluations to support valid inferences about training effects, we probably should try something different. Alternatives to traditional methods for evaluating new Army training are therefore described. Appendixes A through H provide elaboration of evaluation designs and methods presented earlier in the book.

14. SUBJECT TERMS

training evaluation, field trials, analytic evaluations, Army training, statistical power, reliability of scores, validity of scores, validity of inferences

15. NUMBER OF PAGES

16. PRICE CODE

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

THE ELEMENTS OF *TRAINING EVALUATION*

With Chapters on Rationalizations, Elementary Rules of Design and Analysis, Ratings, and Suggestions

BY

JOHN A. BOLDOVICI

DAVID W. BESSEMER

AMY E. BOLTON

THE US ARMY RESEARCH INSTITUTE
for the
Behavioral and Social Sciences
Alexandria, Virginia
2002

Printed 2002, by THE UNITED STATES ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES, a Field Operating Agency under the Jurisdiction of the Deputy Chief of Staff for Personnel.

The findings and opinions in this book are the authors' and are not to be construed as official Department of the Army positions unless so designated by other authorized documents.

ARI has made primary distribution of this book. Address correspondence regarding distribution to

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

Foreword

High costs, environmental concerns, new equipment capabilities that exceed range limits, and political factors in a Post Cold War world have resulted in reductions in U.S. Army field training. At the same time, improvements in simulation technologies have enabled the development of networked simulations that seem to provide a reasonable substitute for collective training in the field. Evaluation of the training effectiveness of these training systems is required by regulation and common sense. In most cases evaluation has involved a comparison of the proficiency of soldiers or units using conventional or field training versus the proficiency of those using the new training approach. While this sounds like a simple comparison, in reality there are significant design, statistical, and cost factors which limit the scope and inferences that can be drawn from most training effectiveness evaluations.

Previous U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) publications have described the design problems inherent in evaluating the training effectiveness of new training systems. These publications also suggested new approaches for evaluation design and analysis of data. These early reports led the U.S. Army Training and Doctrine Command (TRADOC) to sponsor a study that would recommend valid techniques for conducting training effectiveness analyses. The outcome of this study was draft chapters of this book. The study was overseen by a Study Advisory Group (SAG) consisting of representatives from TRADOC's Deputy Chief of Staff for Training (DCST), TRADOC Analysis Command (TRAC), Army Test and Evaluation Command (ATEC), Army Research Laboratory's Human Research and Engineering Directorate (HRED), and the Defense Manpower Data Center (DMDC). Dr. Diana Tierney, DCST served as the Chair of the SAG. The SAG members continued to play an important role in development of the book after the study was formally concluded. Each SAG member reviewed the book and provided comments.

This book represents research conducted by the ARI Simulator Systems Research Unit, whose mission it is to improve the effectiveness of training simulators and simulations. The idea for the book and an early outline were developed as part of the ARI Fiscal Year 1998 Study Program. Further research and actual writing of the book were supported by Work Package SIMTRAIN, Technologies To Enhance Training And Performance Across Simulation Domains.

Zita M. Simutis

ZITA M. SIMUTIS
Technical Director

ACKNOWLEDGMENTS

This book is the result of a project sponsored by Headquarters, US Army Training and Doctrine Command (HQ TRADOC) Deputy Chief of Staff for Training (DCST). The project was the outgrowth of thoughts the authors and Eugenia M. Kolasinski presented earlier about a few elements of training evaluation in a report, *Training research with distributed interactive simulations* and in an article, "How to make decisions about the effectiveness of device-based training." Those earlier publications resulted from work presented to the Office of the Deputy Undersecretary of the Army (Operations Research) as part of financial support for advisory services from COL James E. Shiflett, then US Army Project Manager for Combined Arms Tactical Training. We are grateful for COL Shiflett's monetary support. We are grateful more so for COL Shiflett's apprising us of user perspectives in Army training evaluations, for his encouraging reviews of this and our earlier work, and for promoting and disseminating our work.

Diana Tierney, Research and Studies Branch Chief, Training Development and Analysis Directorate, Futures Training Division, served as TRADOC's technical monitor for the present project. She also was Chairperson of our Special Advisory Group (SAG) and provided several reviews. We are grateful to Dr. Tierney for her thoughtful reviews and good cheer. We thank her especially for unfailing encouragement, bred mainly by unfounded faith that eventually we would get this right.

We thank all the members[1] of our SAG for their technical reviews and for other help and insights:

[1]SAG membership does not signify endorsement of this book.

Contents

SUMMARY xv

INTRODUCTION xxv

I. RATIONALIZATIONS I-1

Rationalizations for Junk Training Evaluations I-3

1. Regulations require demonstrating the effectiveness of new training by estimating transfer of training. This is best done for tactics and maneuver training with large-scale, multi-echelon, combined-arms field trials I-3

2. GAO reviews of training evaluation in the Army call for more testing I-4

3. This is not science; it's just training evaluation I-4

4. We may not have sufficient statistical power to detect significant differences between the scores of compared groups, but our test results will at least put us in the ballpark. Our test is an 80 percent solution I-5

Rationalizations for Junk Reporting I-8

1. These analyses are sometimes done but not reported I-8

2. Some analytic organizations feel presenting such findings may confuse the target audience or distract readers from the purposes of the reports I-8

3. These kinds of findings are considered too technical ("down in the weeds") for the target audience I-9

4. It is common practice not to report these kinds of analyses; not doing so is and has been widely accepted as how we do business I-11

5. Sometimes the data needed to do these types of analyses are not collected in the course of the evaluation I-11

6. Introducing terms such as reliability and validity raises questions and concerns that are not germane to study purposes I-11

Conclusions I-12

II. ELEMENTARY RULES OF DESIGN AND ANALYSIS II-1

1. Consider testing the alternative to the null hypothesis II-1

2. Specify the risk the evaluation customer is willing to take of erroneously detecting no differences between the compared groups' scores II-3

3. Perform power analyses to determine the number of observations necessary to detect differences between the scores of compared groups II-4

4. Increase power by reducing the variability of performance measures within the compared groups II-8

5. Randomly assign soldiers or units to the compared kinds of training II-11

6. Randomize variables whose effects cannot be controlled or measured II-12

7. Establish that the compared groups do not differ in ways that might affect outcomes and alter conclusions..... II-13

8. Equalize or systematically vary the amount of training provided by treatments to prevent confounding with the treatment comparison II-16

9. Allow some time to pass after the end of training before giving posttests, and equalize this time for the compared groups... II-18

10. Use only performance tests whose scores are reliable..... II-19

11. Report test reliabilities and their implications for validity II-22

12. Avoid ceiling and floor effects by adjusting posttest difficulty to produce scores between 75% to 25%, or use more than one posttest with varied levels of difficulty II-24

13. Use conventional analyses of raw scores to estimate training effects II-28

14. Perform separate analyses of training-sensitive and training-insensitive test items.. II-28

15. Interpret null results in terms of confidence intervals and power analyses II-29

16. Address the generality of evaluation results and of attendant inferences II-32

17. Never accept evaluation plans or results at face value II-35

18. Monitor indicators of the value of new training during fielding and implementation over the long term to insure sustained training effectiveness II-35

Conclusion and Recommendation II-36

III. RATINGS III-1

Scope III-1

Misconceptions About Ratings III-2

Advantages of Ratings III-4

Disadvantages of Ratings III-5

Essential Properties of Ratings III-6

1. Estimate and report inter-rater reliability and its implications for validity III-6

2. Estimate and report generality III-7

Designing Ratings for Reliability III-7

Phase I: Rater Preparation III-7

3. Be specific in instructions to raters III-8

4. Provide instructions early enough to allow practice, feedback, and learning III-8

5. Provide practice in observing and rating III-8

6. Test raters III-8

Phase II: Observation III-8

7. Deconstruct multi-dimensional criteria..... III-9

8. Deconstruct multi-dimensional events III-9

9. Make transient events stable III-9

10. Avoid noise in rated events..... III-10

11. Strive for observability in rated events. III-10

12. Require comparative rather than absolute judgments III-10

13. Alert raters to likely errors III-11

14. Allow raters to observe and rate more than once III-11

15. Provide scoring aids or templates..... III-11

16. Do not require raters to process results..... III-12

Phase III: Recording III-12

17. Keep the time short between observing and recording III-12

18. Keep the rating forms simple..... III-13

Additional Sources III-13

IV. SUGGESTIONS IV-1

1. We probably should try something different.. IV-1

2. Meeting the training-evaluation challenge is in some ways analogous to successfully conducting a hasty attack IV-1

3. The complexity of higher-echelon, device-based training guarantees that any single index of effectiveness will be meaningless ... IV-2

4. Higher-echelon, device-based training must be evaluated in systems terms IV-4

5. Approach evaluation of modern, device-based, higher-echelon military training as part of a larger evaluation program applied to the total Army training system and directed toward continual training improvements IV-4

6. Evaluation policies and processes for higher-echelon, device-based Army training programs must be planned as continual, institutionalized parts of a superordinate, total-Army, TQM-like system such as mentioned earlier..... IV-5

The Arsenal..... IV-6

1. Conduct in-device learning experiments to examine the effects of altering training conditions IV-7

2. Conduct quasi-transfer experiments to supplement, replace, or simulate field transfer experiments IV-10

3. Conduct correlational research with archived data IV-13

4. Use efficient experimental designs to control sources of variation and thereby increase statistical power..... IV-14

5. Use quasi-experimental design methods when controlled experiments with randomization are impractical or inefficient. IV-25

6. Evaluate training device capabilities analytically in all phase of the system life-cycle IV-26

7. Improve methods for documenting training by establishing one Army agency with adequate resources to perform the training analysis mission IV-28

REFERENCES REF-1

APPENDIX A
The determinants of statistical power A-1

APPENDIX B
Summary of Cook & Campbell's (1979) Methods for Reducing Within-Group Variance B-1

APPENDIX C
Scratch-Pad Estimates of Reliability and Validity C-1

APPENDIX D
Repeated-Measure Latin Squares D-1

APPENDIX E
Transfer-Efficiency and Savings Estimates E-1

APPENDIX F
ANCOVA: Additional Considerations, Utility for Evaluating Simulation Training F-1

APPENDIX G
The Quasi-Experiment: Estimating Transfer of New Training (e.g., CCTT) to Units' Field Performance G-1

APPENDIX H
Analytic Methods: Three Examples H-1

Summary

Most of this book addresses characteristics of field trials[1] and characteristics of ratings necessary for making valid inferences about training effects, training capabilities, and proficiency. Alternatives to traditional training-evaluation methods also are presented.

The introduction contains our views on the following:

1. **Training evaluation concepts and practices used by the US Army.**
2. **The importance of making valid inferences from training evaluations.**
3. **Common evaluation flaws that threaten our ability to make valid inferences, with emphasis on Type II error (e.g., erroneously concluding that new and old training are equally effective).**

The introduction concludes with rationales for writing this book and our purposes for doing so.

Chapter I describes and rebuts common rationalizations for conducting junk training evaluations. We use "junk" here as in junk science,[2] as a modifier for any evaluations, including many field trials and rating-based evaluations, that permit no valid inferences about training effects. In Chapter I we also present and counter common rationalizations

[1]Field trials refer to training evaluations in which the criterion tests are conducted "on the ground," that is, in so-called field settings. An example is any evaluation in which groups are compared in terms of their scores from Field Training Exercises.

[2]See, for example, Cohn (1994) and www.junkscience.com

for evaluation-reporting practices that preclude estimating the extent to which evaluation findings permit valid inferences about training effects.

Chapter II begins the how-to part of the book[3] with elementary rules of evaluation design and analysis. These rules apply for the most part to the design of field trials and to the analysis and interpretation of data from field trials. The rules are summarized in Table S-1, which also shows likely consequences of failure to apply each rule.[4]

Chapter III deals with advantages and disadvantages of ratings and with rules for their use. The kinds of ratings addressed are those used in the US military for estimating the training capabilities of new training and for individual and collective performance appraisal. The rating rules describe ways to elicit reliable and therefore potentially valid ratings from which valid inferences may be made about the effects of training. A summary of the rating rules is in Table S-2.

[3]Readers interested mainly in evaluation methods, or in the evaluation of evaluation methods, may want to ignore or gloss over the material through Chapter I.
[4]We suggest using Table S-1 as a checklist for estimating the extent to which field-trial plans or results allow valid inferences about training effects. Table S-2 applies similarly to ratings.

Table S-1.
Elementary Rules and Reasons

Elementary Rules	Reasons
1. Consider testing the alternative to the null hypothesis of equality.	Testing the hypothesis that one kind of training is superior to another by a stated amount, rather than testing the null hypothesis of equality, reduces our chance of making a Type II error, that is, of erroneously concluding the compared kinds of training are equally effective.
2. Specify the risk the evaluation customer is willing to take of erroneously detecting no differences between the compared groups' scores.	Knowing the evaluation proponent's willingness to make a Type II error, ß, increases our chances of making informed training decisions in the light of null results (no statistically significant differences between the scores of compared groups). With ß = .20, for example, our willingness to change training policies and practices in light of null results will be greater than with ß = .80.
3. Perform power analyses to determine the number of observations necessary to detect differences between the scores of compared groups.	Power analyses allow us to determine whether our evaluation had enough observations, n, to detect statistically significant differences between our compared groups' scores. Knowing the power of a test is essential for valid inferences about whether null results were due to the absence of differential training effects or to the use of a test that was incapable of detecting the differential training effects.
4. Increase power by reducing the variability of performance measures within the compared groups.	Increasing the statistical power of a test does not necessarily require increasing n. Decreasing the variance of scores within compared groups (methods are Chapter II and Appendix A) also results in increased power and reduces our chance of making a Type II error.

Table S-1 Cont'd
Elementary Rules and Reasons

Elementary Rules	Reasons
5. Randomly assign soldiers or units to the compared kinds of training.	The legitimacy of using conventional (parametric) methods of statistical analysis rests on random assignment of units to the treatments – kinds of training in our case – under examination. Without random assignment, the treatment effects are likely to be confounded by differences between the compared groups that had nothing to do with the kinds of training we wish to compare.
6. Randomize variables whose effects cannot be controlled or measured.	Randomizing variables that we cannot measure or control decreases their effects on evaluation outcomes; without randomization such variables will have unknown and possibly greater effects on evaluation outcomes than the training effects we wish to establish.
7. Establish that the compared groups do not differ in ways that might affect outcomes and alter conclusions.	The results of tests given before the training in an evaluation begins ("pretests") help establish whether the compared groups differed in proficiency or in other ways that might confound evaluation outcomes. Without knowledge of these differences, we cannot determine whether evaluation results were due to the compared kinds of training or to pre-existing differences between the compared groups.
8. Equalize or systematically vary the amount of training provided by treatments to prevent confounding with the treatment comparison.	Amounts of training exert strong effects, perhaps stronger effects than kinds of training, on proficiency: Within broad limits, more training is better than less. Training evaluations that do not equalize or systematically vary amounts of training therefore leave us wondering whether any observed effects were due to the compared kinds of training or to differences between their amounts.

Table S-1 Cont'd
Elementary Rules and Reasons

Elementary Rules	Reasons
9. Allow some time to pass after the end of training before giving posttests, and equalize this time for the compared groups.	Scores from tests given immediately after training are not reliable predictors of future performance, and tests given at various intervals after training will yield various results. The reliability and validity of scores will increase with the use of multiple posttests, as will the validity of inferences we make from the scores.
10. Use only performance tests whose scores are highly reliable.	Reliable test scores are essential for adequate statistical power, for making valid inferences from training-evaluation results, and for decreasing our chances of making Type II errors.
11. Report test reliabilities and their implications for validity.	With test reliability unreported, readers have no objective means for estimating the validity of inferences from evaluation results.
12. Avoid ceiling and floor effects by adjusting posttest difficulty to produce scores between 75% to 25%, or use more than one posttest with varied levels of difficulty.	Ceiling effects mask differences between the scores of compared groups, either of which may have scored higher on a more difficult test. Floor effects mask differences between the scores of compared groups, either of which may have scored higher on a less difficult test.
13. Use conventional analyses of raw scores to estimate training effects. (Avoid using transfer formulas, correlation, efficiency and savings for estimating transfer.)	Transfer formulas, correlation, efficiency and savings estimates do not establish the causal link between training and the outcomes we measured. Conventional analyses, such as analyses of variance and *t*-tests, yield estimates of the extent to which the outcomes we measured could have occurred by chance.

Table S-1 Cont'd
Elementary Rules and Reasons

Elementary Rules	Reasons
14. Perform separate analyses of training-sensitive and training-insensitive test items.	Averaging test scores from training-sensitive and training-insensitive test items diminishes training effects by definition and biases evaluation results in favor of null results and Type II error.
15. Interpret null results in terms of confidence intervals and power analyses.	Interpreting evaluation results in terms of confidence intervals and power analyses is the best way to ascertain whether null results were due to equal effectiveness of compared training regimens or to deficiencies in evaluation design.
16. Address the generality of evaluation results and of attendant inferences.	Any evaluation result is a point estimate, that is, one of a theoretically infinite number of such estimates. It makes little sense to base training policy on any point estimate without knowing the extent to which the estimate represents the theoretically infinite distribution of such estimates.
17. Never accept evaluation plans or results at face value.	Almost all training-evaluation plans will have flaws that severely limit or entirely preclude valid inferences about training effects. Not all evaluators report the gamut of evaluation-design deficiencies that may have caused their results. Common deficiencies may be inferred from the rules and reasons in this table.
18. Monitor indicators of the value of new training during fielding and implementation over the long term to insure sustained training effectiveness.	Repeated training evaluations done over the long term yield inferences about training effects that are likely to be more valid than inferences from any single evaluation. The results of a single evaluation done immediately after training are unlikely to represent the performance of the same soldiers and units tested at later dates.

Table S-2
Rules for Ratings

1. Estimate and report inter-rater reliability and its implications for validity.
2. Estimate and report generalizability.
3. Be specific in instructions to raters.
4. Provide instructions early enough to allow practice, feedback, and learning.
5. Provide practice in observing and rating.
6. Test raters.
7. Deconstruct multi-dimensional criteria.
8. Deconstruct multi-dimensional events.
9. Make transient events stable.
10. Avoid noise in rated events.
11. Strive for observability in rated events.
12. Require comparative rather than absolute judgments.
13. Alert raters to likely errors.
14. Allow raters to observe and rate more than once.
15. Provide scoring aids or templates.
16. Do not require raters to process results.
17. Keep the time short between observing and recording.
18. Keep rating forms simple.

Note: Reliability, statistical validity, and inferential validity likely increase variously with extent to which Rating Rules 3 through 18 are used.

In Chapter IV we suggest that, in light of the consistent failure of Army training evaluations to support valid inferences about training effects, we probably should try something different. Alternatives to traditional methods for evaluating new Army training are therefore described. The alternative methods are:

1. In-device learning experiments.
2. Quasi-transfer experiments.
3. Correlation research with archived data.
4. Efficient experimental designs.
5. Quasi-experimental designs.
6. Analytic evaluations.
7. Improved methods for documenting training.

Appendixes address the following topics:

A. The determinants of statistical power.
B. Summary of Cook & Campbell's (1979) methods for reducing within-group variance.
C. Scratch-pad estimates of reliability and validity.
D. Repeated-measure Latin squares.
E. Transfer-efficiency and savings estimates.
F. ANCOVA: Additional considerations, utility for evaluating simulator training.
G. The quasi-experiment: Estimating transfer of new training (e.g., CCTT) to units' field performance.
H. Analytic methods: Three examples.

The Elements of Training Evaluation

Introduction

Our intent in writing this book is to assist senior commanders and other officials who exercise approval authority over plans for training evaluations. We hope what we have written will increase the willingness and the ability of these individuals to think in terms of the validity or the "invalidity" of inferences from evaluations of new training, and thus to probe for relevant missing information, to assess the strengths and weaknesses of proposed evaluation methods, and to make informed decisions based on elementary principles of evaluation design and analysis and a few simple rules of valid inference. We hope too that reading this book will help military and civilian leaders who use evaluation results in making decisions about adopting new training: The same principles that guide planning also can be used to assess the validity of evaluation results and conclusions. We shall also be pleased if the views we present impart a reminder or two to the civilian evaluators upon whom the Army relies for designing, analyzing, and making valid inferences from training evaluations. And if readers note some relevance of what we have to say for evaluations other than military training, that will please us too.

Expenditures and Evaluations

The US Army allocates millions of dollars and large amounts of personnel time to training that prepares individuals and units for actions in war and other military operations. Large capital investments are especially being made in simulators and other training devices that support training in virtual and live environments. Justifying these expenditures of money and manpower requires describing what training produces in terms that are understandable to all military constituencies: executive, legislative,

and public. Training evaluation is the process for developing objective descriptions of training results.

Evaluation Purposes

Cronbach (1969), Patrick (1992), Boldovici and Bessemer (1994), and others have enumerated various training-evaluation purposes, which include program improvement, administrative and organizational decisions, compliance with regulations, and decisions about trainees. Achieving any evaluation purpose rests on the validity of the evaluation result and the validity of inferences we make from the evaluation result. We imply the obvious here: An evaluation result can be valid or invalid, as can be the inferences we make from the evaluation result irrespective of whether the evaluation result is valid or invalid. A less obvious implication is that training evaluators do not usually report estimates of the validity of their evaluation results,[1] and we therefore have no objective grounds for estimating the validity of any inferences made from evaluation results. We resist here the temptation to speculate about reasons for this state of affairs and trust readers will draw their own conclusions from Chapter I.

The Notion of Valid Inference

The validity of an inference is never unequivocally established. Whether an inference is invalid on the other hand is easily established, especially when the inference is from a training evaluation or from other evaluations. Estimating the validity of an inference is a simple three-step process, the elements and computational procedures for which are discussed later in this book.[2] The first step in the validity-estimation process is to estimate the reliability, that

[1]We know of no exceptions in evaluations of Army training.
[2]See Appendix B for examples.

is, consistency, of the scores obtained in an evaluation. Reliability is typically given as a range from 0 to 1.0, with 0 indicating no reliability and 1.0 indicating perfect reliability. Reliability as low as .50 may be acceptable for training evaluations with large samples. For test scores that feed decisions about particular individuals or units, however, reliability should be .90 or greater.

The second step in the validity-estimation process is to use the reliability estimate to estimate the scores' validity, variously defined, but for our purposes mainly the predictive value of scores, for example, the ability of gunnery-simulator scores to predict live-fire scores or the ability of field-exercise scores to predict Combat-Training-Center scores. For validity defined in that sense, the axiom is that the validity of scores cannot exceed the square root of the reliability estimate.[3] Note that the validity estimate is an estimate, not of a guaranteed validity associated with any given reliability, but of the maximum validity that is possible with any given reliability. Acceptable reliability thus never guarantees acceptable validity. But unacceptable reliability guarantees unacceptable validity.

The third step in estimating the validity of inferences from training evaluations does not require arithmetic. What is required is a moment's rational thought aimed at answering the following question: Given the estimated reliability of scores and the maximum possible validity of scores calculated from the reliability estimate, what are the chances that inferences from the scores, for example, inferences about the relative effects of old and new training, will be valid? With reliability in the nineties, for example, and maximum validity by definition in the even

[3]Because the reliability of scores is given in decimal fractions and the validity of scores cannot exceed the square root of reliability, maximum validity will be greater than reliability – except, of course, in the unlikely case reliability is 1.0.

higher nineties, we have a prayer – but no guarantee – of making valid inferences from the evaluation results. With reliability in the twenties, for example, and maximum validity by the square-root rule in the fifties,[4] we should view with extreme caution, that is, doubt, inferences from any evaluation results.

Without an estimate of the reliability of scores from an evaluation, we have no grounds for estimating the validity of those scores. And without estimates of the validity of the scores, we have no objective grounds for estimating the validity of inferences from the evaluation. How then may we estimate the validity of inferences from the scores such an evaluation produces? The answer is, Without objectivity. David Hume's advice comes to mind:

> If we take in our hand any volume . . . let us ask, *Does it contain any abstract reasoning concerning quantity or number?* N . . Commit it then to the flames, for it can contain nothing but sophistry and illusion (Hume, *ca.* 1760).

The Concept of Evaluation

Evaluating Army training is conceptually a simple matter, involving, for example, (a) empirical comparisons between the proficiency of soldiers or units using conventional or field training and the proficiency of soldiers or units using new (usually device-based) training, (b) expert ratings of program

[4]Low reliability means that much of the variation in our scores is random, that is, not caused by the compared kinds of training. The small proportion of non-random, training-induced, variation suggests *some* validity. But we are left not knowing which parts of the score training affected and therefore not knowing what the score means.

effectiveness or training capabilities,[5] or both (a) and (b). In practice, however, this conceptually simple matter presents technical and logistic threats that always limit and often preclude valid inferences about the result, that is, the effect,[6] of training. Failure to understand threats to valid inference leads to accepting erroneous conclusions about military training effects and provides the foundation for a house of cards on which to build inappropriate kinds and amounts of training with concomitant threats to readiness, the national defense, and the longevity of our progeny in armed conflict. Our security depends on understanding and countering threats to valid inferences about military training effects.

The Practice of Evaluation

Tests of new Army training typically compare the effects of conventional or field training to an altered training regimen in which part of the conventional or field training is replaced by simulator- or device-based training.[7] Because device-based training may be proposed to replace parts of conventional training, a question naturally arises about whether the proposed substitution will adversely affect soldiers' proficiency as compared to the proficiency

---

[5]Training-capabilities evaluations also are known as analytic evaluations. Examples are in Chapter IV, Rule 6, and in Appendix G.

[6]We shall try, perhaps unsuccessfully, to avoid indiscriminate use of the term "training effectiveness." Our reasons for wanting to do so include the imprecision of the term and its connotation of a unitary metric. Abstractions such as training effectiveness ". . . are at best distracting and at worst misleading. They are distracting because careful readers are forced to translate . . . 'effectiveness' into the operations that were actually performed. The abstractions are misleading because less careful readers may not go through any translation process at all: The 'effectiveness' of training devices [for example] becomes incorporated as a fact with little or no understanding of what caused what" (Boldovici, 1987, p. 258).

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

of soldiers who train with existing, conventional means. That question easily translates to a null hypothesis of equality of training effects, H_o: $M_c = M_d$, where M_c and M_d are the mean scores of the conventional and device groups, and may be formulated as such by training evaluators. Thus is laid the foundation for the house of cards.

The House of Cards, Part I

As sure as questions about the substitutability of new training for parts of old training lead to testing null hypotheses about the equality of new and old training, hypotheses about the equality of new and old training lead to using two-group, *t*-test evaluation designs: One group receives the old training, another group receives the new training, both groups take identical tests to assess their proficiency after training, and analyses are performed to estimate the extent to which the difference between the compared groups' average scores could have happened by chance. This paradigm is commonly used in evaluations of new Army training.[8] Using such a paradigm makes intuitive sense at some level: Brush with Regular Crest, brush with Flouristanated Crest, and see which brushers wind up with fewer cavities. But as a decision- or inference-support tool for military training, the two-group, new-vs.-old, *t*-test design is inadequate; its use leads, at best, to no legitimate inferences about training effects and, at worst, to invalid inferences about training effects. Invalid inferences about military training effects are at least as misleading and therefore as dangerous as invalid inferences about weapons effects.

As an example of the dangers inherent in invalid inferences from military training evaluations, consider the following: Two-group, new-vs.-old-

[8]See Boldovici and Bessemer (1994) for a review of examples.

training evaluations are conducted, and analysts find no statistically significant differences between the post-training test scores of the compared groups. Prominent evaluators and military leaders review the results of such evaluations and conclude that because no statistically significant differences between the post-training scores of the two groups were found, no such differences exist, and the new and the old training must therefore be equally effective.[9] If we believe that the new training and the old training are equally effective, then we easily make the next inductive leap, namely, that substituting the new training for the old poses no risk. This line of thinking is so fraught with invalid inferences and is so pernicious – in terms of what may happen as the result of dead-wrong inferences about the substitutability or equality of new training for old training – that we yield here to the conceit of reproducing an earlier-published rebuttal.

The Myth of Equal Effectiveness[10]

Statisticians, biomedical researchers, and behavioral scientists have publicized errors in examinations of the differential effects of two or more treatments [new vs. old training in the present case]. The publicity about those errors seems to have been ignored by many applied behavioral researchers, including some responsible for evaluations of new, usually device-based training in the US Army. Ignoring the causes and effects of the common evaluation errors, and especially errors associated with hypotheses of equal effectiveness of conventional training and device-based training, leads to logical contradictions, threats to readiness, and no statistically legitimate ways to examine the effects of OPTEMPO[11] alterations.

---

[9]See Orlansky (1985) and Wickham (1983) for examples.
[10]From Boldovici and Kolasinski (1997, pp. 123-125).
[11]OPTEMPO is an abbreviation for operating tempo; it refers to "the annual operating miles or hours for the major equipment

Our chief concern is with the belief that null results, that is, finding no statistically significant differences between compared groups' scores as the result of, for example, new training vs. old training, signify equal effectiveness of the compared kinds of training.

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

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

---

system in a battalion-level or equivalent organization" (National Simulation Center, 1994). OPTEMPO also is, as Robin Rose (personal communication, December 2000) noted, ". . . a DOD term, not just an army term, see for example http://www.defenselink.mil/news/Aug1999/n0818199_9908181.html Also there is a second sense in which OPTEMPO refers to the cost or budget items associated with the vehicle miles traveled (see for example http://www-cgsc.army.mil/cdd/f545/f545-no.htm#OPTEMPO)."

differences visible to the naked eye and (2) by misinterpreting null results.[12]

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

In addition to providing precedent for spuriously demonstrating the equivalence of training and no training, the myth of equal effectiveness of field training and device training paves the way for closing training and maneuver areas and for additional decreases in resources that attend field

---

[12]A reviewer noted,

> [If] . . . field training is better than device training for some tasks and worse for others . . . then a unitary comparison between the training methods makes little sense. We should be trying to allocate tasks to training methods. Perhaps a third way to produce equally effective training methods, then, would be to sample tasks that favor field training and tasks that favor device training in roughly equal proportions.

We agree on all counts.

training. Downsizers' contentions are easy to foresee: "If device training and field training are equally effective, then what harm can come from additional substitutions of device training for field training, that is, from additional reductions in OPTEMPO?" The flaws in that line of thinking can be exposed by applying legitimate methods for examining the equivalence (and non-equivalence) of alternative kinds of training—methods which we shall discuss shortly and which, to the best of our knowledge, have not been used in evaluations of device-based training in the Army. Military leaders and the device evaluators who advise military leaders need to understand the differences between legitimate and illegitimate methods for establishing the equivalence of alternative kinds of training. That understanding is essential to ensuring the use of legitimate methods for examining the effects of device-based training and of OPTEMPO alterations.

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

The House of Cards, Part II

Various flaws threaten the validity of inferences that are or can be made from training-evaluation results. The threats to valid inference, in a rough chronology of the order in which they usually occur, include:

1. Flaws in framing evaluation questions (addressed above).
2. Flaws in designing evaluations (addressed above).
3. Flaws in executing evaluations.
4. Flaws in analyzing evaluation results.
5. Flaws in reporting evaluation results.
6. Flaws in interpreting evaluation results.

These flaws, whose avoidance and correction are addressed throughout this book, characterize evaluations of new Army training. For samples of the flaws listed above, let us consider flaws in a hypothetical example of an Army evaluation of new training.[13]

1. Flaws in framing evaluation questions.

Army evaluations, as discussed above, often address the question of equality of old training compared to new training. Questions of equality of training effects are misleading for reasons in addition to the flaws mentioned earlier: The absence of differences, that is, equality of training effects, cannot be proved.

2. Flaws in designing evaluations.

To address the question of equality between new training vs. old, evaluators of new Army training usually use the two-group, *t*-test design discussed above: The average test scores of soldiers or units

[13]Our hypothetical example comprises an amalgam of flaws in evaluations of new Army training, particularly SIMNET and more recently the Independent Operational Test and Evaluation (IOT&E) for CCTT.

who receive the new training are compared to the average test scores of soldiers or units who receive the old training. The chief flaw here is that the design contains no control group.[14] Assuming that the post-training proficiency of either group was caused by the training they received is therefore unwarranted by any result. The most we can discover from implementing such a design is that the new training was less or more effective than old training of unknown effectiveness. Such results are irrelevant for decisions about whether to adopt new training and for decisions about how to improve existing training.

3. Flaws in executing evaluations.

A common flaw in executing military training evaluations involves failure to control for amounts of training.[15] Amounts of training are not typically reported in evaluations of new Army training. Suppose the evaluation shows the new training to be more productive than the old training. As with fishing-lure infomercials, we should wonder whether the superior performance of the group using the new lure was due to the new lure's superiority to old lures or to the amounts of time the compared groups had their lines in the water – or, in the present case, time spent training.[16]

4. Flaws in analyzing evaluation results.

Evaluators of new Army training do not typically report estimates of the reliability of the test scores obtained by their compared groups. Without estimates of reliability, we have no grounds for

---

[14]A control group is one that is treated identically to the new-training group except that it receives no training. Using a control group as defined here is wholly feasible for evaluations of sustainment training: There is no danger in using a no-sustainment-training control group in evaluations of new sustainment training. This point is elaborated in Chapter II.

[15]This flaw, also known as "confounding," is elaborated later.

[16]Krueger demonstrated that more training was better than less in 1929.

estimating the validity of results or the validity of inferences we may make from results. Estimates of the generality of results also are never reported; officials with decision-making responsibility therefore have to make inferences based on results whose generality is not known. The results of power analyses or of confidence intervals, which are essential for determining the cause of a finding of no difference between the effects of new training and old, are never reported. Decision makers and other readers therefore have no grounds for judging the validity of inferences they or evaluators make from evaluation results.

5. Flaws in reporting evaluation results.
The ubiquity of evaluation flaws notwithstanding, evaluators of new Army training do not typically report those flaws. Nor do they typically report the certain effect of those flaws on inferences from evaluation results, namely that valid inferences from the evaluation results are not possible.

Evaluation flaws are more likely to be reported in Army training evaluations than are the effects of evaluation flaws. We hope this subtlety warrants the following digression: The cover letter for the CCTT Independent Operational Test and Evaluation (Director, Operational Test and Evaluation, 1998), for example, says, "Due to time and resource constraints, the treatment unit for IOT consisted of a single battalion task force . . . ." We guess lay readers, including the members of Congress to whom the cover letter and report are addressed, gloss over that *mea culpa*, perhaps in the conventional belief that evaluation results with inadequate sample sizes have some systematic relation to results with adequate sample sizes. That belief is not warranted because (a) results that would be got from evaluations with adequate samples cannot be predicted from results got with small samples, and (b) the use of inadequate sample

sizes so compromises training evaluations that detecting training effects that exist is not possible.[17]

The author of the cover letter wrote nothing to inform readers of errors in the conventional wisdom regarding his small-sample deficiency,[18] perhaps because he too believes the conventional wisdom.[19]

6. Flaws in interpreting evaluation results.
The chief flaw in interpreting results of Army training evaluation is in declaring the new and the old training equally effective.[20] This flaw results from a variety of factors, of which insensitive measurement is an example. (On a scale graduated in tons, a feather and a piano are equally heavy.) Failure to separate an evaluation finding from one's inference about the evaluation finding is, in our view, the ultimate conceit: (a) I found no needle in the haystack, therefore there is no needle in the haystack. (b) I found no differential training effect, therefore there is no differential training effect.

Valid inferences about training effects from the results of flawed training evaluations are hard to imagine. Valid inferences from such evaluations are, in fact, impossible. Objectivity about popular inferences from flawed Army evaluations of new training is what this book is about.

---

[17]This is the statistical-power problem. Its solution is in Chapter II.
[18]See summary of Tversky and Kahneman's (1971) treatise on belief in the non-existent law of small numbers and our discussion in Chapter II.
[19]The writer of the cover letter did, however, complete his *mea culpa* by saying, "This sample is small and highly correlated for supporting inferences about training effectiveness." We have no idea what that means.
[20]See earlier treatments in House of Cards, Part I, and Orlansky (1985) and Wickham (1983).

Training Evaluation in the US Army

Training evaluation in the US Army comprises field trials and ratings. This is a rough distinction at best, inasmuch as ratings often are used to generate scores in field trials. Field trials such as Concept Evaluation Program Tests, Force Development Tests, and IOT&E usually are used for establishing summative training effectiveness estimates,[21] rather than for formative or diagnostic purposes. Field trials of Army training typically use, as noted earlier, the two-group, *t*-test design that compares transfer of training resulting from conventional, existing training to transfer resulting from an altered training regimen in which part of the existing training is replaced by new training. Current new training in the Army is based largely on simulators and other training devices. Examples are evaluations of SIMNET and the IOT&E for the Close Combat Tactical Trainer (CCTT).[22]

Ratings are used, as noted above, in some parts of field trials for scoring soldiers' or units' proficiency. Examples of such use include ratings of maneuver and tactics. In addition to their use in field trials for estimating new-training effectiveness, ratings also are used for diagnostic purposes – diagnostic not only of soldiers' or units' performance, as in after-action reviews at the Combat Training Centers, but

[21]Regarding summative training evaluations in the Army, Diana Tierney (personal communication, December 2000) noted, "They are done very infrequently but probably should not be done at all – and certainly should not be held up as the superior way of evaluating training (at least not the way this has to be implemented in the real world Army)." We agree.

[22]Most of our evaluation experience in the Army is with simulator- or device-based training, most recently the CCTT, and less recently the CCTT's predecessor, SIMNET. The examples we use throughout this book therefore often refer to evaluations of training that uses CCTT or SIMNET. We do not present deficiencies in those evaluations as unique. They just happen to be examples we know something about.

diagnostic also of new-training, and especially device-training, capabilities.[23]

Rationale for Rationales

During the planning and early drafting of this book, we decided not to present rationales for identifying, avoiding, and countering threats to valid inferences from training evaluations. Our thinking was driven by various motives. One motive involved self-evidence: Military leaders and their civilian advisors need no more to be persuaded of the consequences of invalid inferences from training evaluations than they need to be persuaded of the consequences of invalid inferences from weapons-systems evaluations.

Our motives also included wanting to be brief and wanting to emulate the spirit of Strunk and White (1979), who presented various rules for using the English language in ways that make sense, but presented few if any rationales for doing so.[24] In addition to wanting to be brief, Strunk and White seem to have decided, for reasons we have no hope of knowing, but suspect had something to do with self-evidence, that they would not burden readers with an enumeration of reasons for avoiding constructions such as, "The cows has come home to roost." Our thinking in turn was that our readers would need no more reason for avoiding threats to valid inferences from Army training evaluations than Stunk and White's readers would need for seeking alternatives to the cows has come home to roost. Discussions with persons associated with military

[23]Examples of ratings used to diagnose new training capabilities include the work of Drucker and Campshure (1990) with SIMNET, of Burnside (1990) with SIMNET, and of Sherikon Corporation (1995) with the CCTT. These analytic evaluations are summarized in Chapter IV, Rule 6, and in Appendix G.

[24]Readers who view Strunk and White's omission of rationales as a deficiency may find remedies in Richard Mitchell's *Less Than Words Can Say: The Underground Grammarian* (1979).

training evaluations proved wrong our early thinking about the self-evidence of rationales for needing valid inferences from military training evaluations. Those discussions were rife with rationalizations for conducting training evaluations that might yield no valid inferences, or misleading inferences, about training effects and training policy. The rationalizations, because of their ubiquity, and the rationalizers, because of their majority, led us to add the following section about rationales, and Chapter I: Rationalizations.

Rationales

The rationales for writing this book can be inferred from our discussion leading to this point: Invalid inferences from training evaluations are logically and rationally untenable by definition. Believing, adopting, and promulgating such inferences pose threats to readiness, the national defense, and the lives of our progeny.

Purpose

Our purpose is to provide readers with objective methods for assessing the utility of plans for, and the validity of inferences from, training evaluations. In the course of so doing, we shall:

1. Summarize and rebut rationalizations for flawed training evaluations and evaluation reporting (Chapter I).
2. Present a few elementary rules of evaluation design and analysis (Chapter II).
3. Discuss consequences of, and alternatives to, flawed training evaluations (Chapter II).
4. Discuss the role of ratings in training evaluations (Chapter III).
5. Recommend new directions and methods for future training evaluation (Chapter IV).

I

Rationalizations

Practical constraints almost always diminish our ability to adhere to elementary rules of evaluation design and analysis. We can barely imagine situations, for example, in which random assignment of military units to experimental and control treatments would be practical, or in which sufficient numbers of battalions would be available to meet statistical power requirements for evaluations with battalion-level field trials, or in which military test organizations would spend much money to improve the reliability of ratings. Against those practical constraints, we hope Army decision makers will weigh the consequences of diminished possibilities for valid inferences about training effectiveness. Those consequences, as noted later, affect readiness, the national defense, and the ability of our progeny to survive armed conflict. It goes without saying that whether to proceed with evaluations so compromised as to preclude valid inferences about training effects is a judgment call. Improving the judgment in the call is one of the things we hope this book will accomplish.

Notwithstanding practical constraints and their effects on valid inferences about training effectiveness, Horst, Tallmadge, and Wood's (1975) contention remains unassailable: Practical constraints may prevent evaluators from doing controlled experiments, "but many problems in current evaluation practices could be avoided with little or no increase in cost or effort" (p. 2). Reasons for not implementing low- or no-cost solutions to problems that threaten valid inferences from training evaluations are hard to fathom. Rationalizations for decisions to implement

evaluations that hold no possibilities for valid training-effectiveness inferences abound however. We present here a few common rationalizations we have heard, not as broadsides against our colleagues in military training evaluation,[1] but for two purposes we hope readers will find useful. The first purpose is in the spirit of, "Ye shall know them by their fruits" (Matthew, 7:16). That is, we have heard these rationalizations so often that we have come to view them as rules of thumb, "red flags," that attend evaluation busts. Such rules of thumb are in our view essential for informed decision making on the part of evaluation-planners' customers.

The second, like our first, purpose for discussing rationalizations also derives from our concern for informed decision making: The rationalizations are just that – attempts "to devise self-satisfying but incorrect reasons for one's behavior" (Houghton Mifflin Company, 1984, p. 976). Decision makers should never accept the rationalizations at first blush, but should question evaluation planners to determine whether low- or no-cost evaluation improvements are available. Many such improvements are described later in this book. If no improvements are forthcoming the choice is easy: Accept or abort.

The rationalizations to which we refer may be heard, not only in defense of compromised training evaluations, but also in defense of reporting that does not allow readers to assess the credibility of

[1]We have been involved in just as many compromised Army training evaluations as the next guys. We have, however, been careful to inform readers of the effects of our compromises on the possibilities for valid inferences. In their report on the widely touted role of SIMNET in the 1987 US win of the Canadian Armor Trophy, for example, Kramer and Bessemer wrote, "It is impossible to determine . . . whether SIMNET training benefited, reduced, or had no effect on the performance of the US platoons in the CAT competition" (1987, p. 28).

inferences from evaluation results. We address those two sets of rationalizations in the following two sections.

Rationalizations for Junk Training Evaluations

"Junk" is used here as in junk science[2] and refers to evaluations whose design, execution, or analysis precludes valid inferences about training effects. Our experience suggests that, despite the impossibility of drawing valid inferences about training effects, many Army decision makers, their civilian counterparts, and their advisors routinely defend plans, administration, and results of junk training evaluations with one or more of the following rationalizations:

1. *Regulations require demonstrating the effectiveness of new training by estimating transfer of training. This is best done for tactics and maneuver training with large-scale, multi-echelon, combined-arms field trials.*

Our review of TRADOC's *Training Development Management, Processes, and Products* (1995) and a companion review by Henry K. Simpson (personal communication, October 1998) uncovered no requirement to demonstrate transfer of new training. The tradition of trying to demonstrate transfer of new Army training rather seems to have begun with operational tests of the Unit Conduct of Fire Trainer[3] and continued through operational tests of SIMNET and CCTT. For a variety of reasons, many of which can be inferred from later discussions in this book, these attempts failed. One reason for the failures is that training evaluators have not sufficient numbers or kinds of countermeasures to overcome the threats to valid inference that inhere in one-shot, two-group,

[2]See Cohn (1994) and www.junkscience.com for examples.
[3]See Kuma and McConville (1982) for an example.

multi-echelon, combined-arms field trials of training effectiveness.

2. *GAO reviews of training evaluation in the Army call for more testing.*

At least three GAO reports (1986, 1990, 1993) address the need for more testing of new Army training. In contrast to their emphasis on amounts of testing, the reports address kinds of testing only tangentially. The reports do not prescribe kinds of tests and evaluations and therefore cannot legitimately be used to justify doing more of what we have done poorly. Attempts to justify more one-shot, large-scale, multi-echelon, combined-arms evaluations of new Army training on grounds of GAO recommendations are gratuitous. A need may well exist, as stated in the GAO reports, for more testing. The greater need is, however, for better testing – that is, for testing more likely than its predecessors to permit valid inferences about training effects.

3. *This is not science; it's just training evaluation.*

This rationalization has the potential to rise to the level of word-smithing. But to do so it would have to make more sense. Inferences do not care whether they are from science or from a training evaluation. Inferences either are valid or they are not. It matters not whether we call the set of operations whence inferences come science or training evaluation. Renaming the operations does not change the rules of inference.

As for the modifier "just," we are hard-pressed to fathom the depth of muddled thinking underlying its use. Does "just" mean "only," with attendant implications of "merely" and "unimportant?" Or does "just" mean training evaluation is just one more ticket-punching inconvenience for tourists on the Army's training-acquisition express? One hopes the

purveyors of such an outrageous rationalization are just unwilling to consider, rather than just incapable of considering, the consequences of just making erroneous inferences from military training evaluations.

4. *We may not have sufficient statistical power to detect significant differences between the scores of compared groups, but our test results will at least put us in the ballpark. Our test is an 80 percent solution.*

This rationalization reflects belief in a law of small numbers; that is, a belief that evaluations and other research with small numbers will yield results that reflect the results we would have got had we used large numbers. Belief in a law of small numbers is a misconception held, not only by persons unschooled in statistics and evaluation, but also by training evaluators, scientists, and other researchers. Tversky and Kahneman addressed this belief in 1971. After more than a quarter century, few persons who plan military training evaluations, and even fewer who use the results of such evaluations, have got Tversky and Kahneman's message. The message is,

> People have erroneous intuitions about the laws of chance. In particular, they regard a sample randomly drawn from a population as highly representative, that is, similar to the population. . . (p. 105).

Because arguments against the belief in laws of small numbers are counterintuitive, we urge readers to study Tversky and Kahneman's tract at leisure.

We also commend to our readers Gawande's article in the 8 February 1999 *New Yorker*. The article is about how Tversky and Kahneman's thinking affected the Center for Disease Control's policies and methods for identifying cancer enclaves. Similarities are numerous

between the decisions faced by CDC decision-makers and by persons responsible for establishing and implementing military training-evaluation policies.

Against the chance a reader or two may ignore our suggestions for recreational reading, we reproduce here Tversky and Kahneman's exposé of the stigmata by which the believer in laws of small numbers betrays himself:

- *He gambles his . . . hypotheses on small samples without realizing that the odds against him are unreasonably high. He overestimates [statistical] power.*

- *He has undue confidence in early trends (e.g., the data of the first few subjects) and in the stability of observed patterns (e.g., the number and identity of significant results). He overestimates [statistical] significance.*

- *In evaluating replications, his or others', he has unreasonably high expectations about the replicability of significant results. He underestimates the breadth of confidence intervals.*

- *He rarely attributes a deviation of results from [his] expectations to sampling variability, because he finds a causal "explanation" for any discrepancy. Thus, he has little opportunity to recognize sampling variation in action. His belief in the law of small numbers, therefore, remains intact* (p. 109).

In addition to flying in the face of basic statistics, the ballpark/80% rationalization is assailable on rational grounds. For openers, we have no *a priori* criteria for judging whether we are in the ballpark, which is an issue of generality of results. Resolution of any generality issue requires replication, and replication is

not feasible for multi-million-dollar tests of new military training. The ballpark, like many so-called 80% solutions,[4] is defined after the fact as wherever the results happen to put us. If we conduct compromised evaluations with new training and find no statistically significant differences between the scores of compared groups (e.g., conventionally trained vs. device trained), then the results are, contrary to the ballpark thinking, no better than guessing: Random or error variance exceeded the variance due to the compared training regimens, and our field trials might as well have not been conducted. That is especially true for cases in which we suspected or knew in advance that the power of our trial was so weak as to preclude finding statistically significant differences between compared groups' scores.

The ballpark line of thinking disconcerts additionally because null results in military training evaluations are readily taken, without supporting analyses, as evidence that conventional training and new training are equally effective. As noted throughout this book, null results in training evaluations can ensue from causes other than equal effectiveness of the compared training. The inductive leap from finding no differences to declaring equal effectiveness[5] is as dangerous in

[4]One wonders what characteristics define 80% solutions. How shall we distinguish 80% solutions from, say, 78% solutions? Or for that matter 100% dead-wrong solutions?

[5]One need not look far to see scientists and science reporters joining the leap. On the front page of an Orlando newspaper is a story by Recer (1999) with this proclamation from Elizabeth Harvey of the University of Massachusetts at Amherst: "There was no difference between children whose mothers were employed vs. children whose mothers were not employed during the first three years. Being employed is not going to harm children." The article is in *Developmental Psychology*, a refereed journal of the American Psychological Association (Harvey, 1999).

military training evaluation as it would be in bioequivalence research[6] – perhaps more so.

Rationalizations for Junk Reporting

In the course of a chat with about a dozen representatives of DoD and US Army training-evaluation organizations, we asked why the discussants and their colleagues did not routinely report the results of (a) power analyses, (b) the reliabilities of obtained scores, (c) the implications of reliability for validity, and (d) the meaning of null results in relation to confidence intervals. Knowledge of these four aspects of training evaluations is necessary for allowing objective readers to estimate the likelihood that an evaluation permits valid inferences about training effects. A summary of the discussants' replies follows, as do our observations on each.

1. *These analyses are sometimes done but not reported.*

This rationalization begs the question, namely, Why do Army training evaluators not routinely report the results of power and other analyses necessary for estimating the validity of their inferences?

2. *Some analytic organizations feel presenting such findings may confuse the target audience or distract readers from the purposes of the reports.*

[6]See Blackwelder (1982) for examples. The danger, of course, is that Type II errors in bioequivalence research lead to life-threatening decisions. Similar effects of Type II errors in military-training evaluations are obvious: "Evaluators [of military training] should be as concerned about errors in examining the equivalence of alternative training regimens as biomedical researchers are about errors in examining the equivalence of alternative pharmacological treatments. In both cases, evaluation results factor into life-or-death decisions" (Boldovici & Kolasinski, 1997, p.125).

"Paternalism . . . *n.* A policy or practice of treating people in a paternal manner, esp. by taking care of their needs without giving them responsibility" (Houghton Mifflin Company, 1984, p. 861). We hope this rationalization represents a misperception on the part of the person who reported it. Misperception or not, readers interested in estimating the validity of inferences from military training evaluations do not need "some analytic organizations" to protect them from their own thinking.

3. *These kinds of findings are considered too technical ("down in the weeds") for the target audience.*

We were unable, thanks to the passive voice in this rationalization, to ascertain who is doing the considering. The considerer seems in any event to underestimate the ability of target audiences – the great majority of whom our experience suggests hold at least bachelor's degrees – to understand elementary arithmetic, logic, and plain English.

Recall that the "kinds of findings" in this rationalization are (a) power analyses, (b) reliability of scores, (c) implications of reliability for validity, and (d) null results in relation to confidence intervals. Those four concepts undoubtedly comprise a few mathematical and logical esoterica that are "too technical" for some members of our college-graduate audiences. And those four concepts just as undoubtedly are explicable in plain English, *QED*:

(a) Power analyses yield probabilities that an effect, for example a difference between two groups' average scores, can be detected if such a difference indeed exists. Power (probability) of 1.00 means chances are good of finding existing differences between compared groups' scores. Power of .00 means

chances are not good. Power of .50 means we should think twice before spending taxpayers' money trying to detect the difference.

(b) Reliability is an estimate of the consistency of scores. Reliability of 1.00 means the scores are wholly consistent; reliability of .00 means the scores are wholly inconsistent. Reliability of .50 means that half of each score is consistent, and half of each score is not consistent.

(c) The main implications of reliability for validity are that low reliability of scores guarantees low validity of scores and inferences, but high reliability does not guarantee high validity. Inferences from reliable scores are therefore more likely, but not guaranteed, to be valid than are inferences from unreliable scores. The main implication of these implications is that without access to reliability estimates we can estimate the validity of inferences from training evaluations less well than we can play baseball without a bat.

(d) Confidence intervals, for the cases discussed in this book, tell us the range of differences between compared groups' scores that our statistical test would call non-significant. Consider the following example: We have a test in which the possible differences between our compared groups' mean scores range from, say, 0 to 50. A training evaluation yields a null result, that is, we find the differences between two compared groups' mean scores to be statistically non-significant. *Never make the inductive leap from a null result to a declaration of equal effectiveness* (of the compared training regimens). *And beware of those who do.* Faced with a null result, ask to see the confidence interval. If the confidence interval is narrow, containing for example, 3 of the 51 possible differences in our example, any inequality between the compared groups' scores is trivial. If on the other hand the confidence interval is wide,

containing say 30 of the 51 possible differences in our example, then the evaluation in question was a bust; for a variety of reasons, most likely inadequate statistical power, the evaluators could not have detected real between-group mean differences even if those differences jumped up and bit them on the elbow.

4. *It is common practice not to report these kinds of analyses; not doing so is and has been widely accepted as how we do business.*

This is true. How we do business in military training evaluation is, however, unproductive and by definition inefficient. Our thoughts on better ways to do business appear later in this book.

5. *Sometimes the data needed to do these types of analyses are not collected in the course of the evaluation.*

The data needed to conduct the four "types of analyses" in question here are, in the case of field trials, the scores of the compared groups, and in the case of ratings, ratings from two or more raters and in some cases from only one. Training evaluations that do not routinely yield one or both of these kinds of data are hard to imagine.

6. *Introducing terms such as reliability and validity raises questions and concerns that are not germane to study purposes.*

We regret failing to probe the thinking behind this absurd rationalization: We find training evaluations for which the concepts of reliability and validity are not germane impossible to imagine.

Conclusions

The rationalizations for junk reporting are noteworthy on at least two counts: (a) any suggestion of personal responsibility on the part of discussants is absent,[7] and (b) the constraints under which the discussants assume training evaluators operate are inconsistent with our experience.[8]

Evaluators who neglect to interpret their results in terms of power analyses, reliability, implications of reliability for validity, and confidence intervals deny readers any chance of estimating the validity of inferences from the evaluations. Such evaluators invite suspicion of playing fast and loose with the data. There is little danger of inviting suspicion, of course, if readers do not know what fast and loose look like. Whether and the extent to which evaluators are obliged to educate their readers is an open question. That question gets closed quickly in our view when it is the readers' money that is paying for the evaluation.

---

[7]One reviewer expressed surprise that no discussant said, "The devil made me do it."

[8]In contrast to the nearly carte-blanche we have enjoyed as evaluation reporters, we can recall only one case in which a supervisor tried to censor discussion of reliability and validity.

II

Elementary Rules of Design and Analysis

1. Consider testing the alternative to the null hypothesis.

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

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

The obvious way to avoid the problem implied by Fisher, that is, erroneous acceptance of H_O,[1] is never to accept H_O and thus avoid the possibility of accepting H_O erroneously. That prescript is logically irrefutable; it is, however, intellectually unsatisfying because null results engender the immediate question, "Did we find no differences because there are no differences or because flaws in our evaluation precluded finding differences?"

A more satisfying way to assess null results than to dismiss all out of hand is to estimate the chances, represented by β, of making a Type II error for the minimal mean difference (effect) that the evaluation customer considers to be important.[2] That difference is the smallest effect, judged by responsible authorities, that indicates the new training is inferior enough to justify abandoning its additional development or use.

There are at least two additional ways to address Fisher's concern about accepting null hypotheses. One is to forgo hypothesis testing and instead use confidence intervals.[3] Another way was suggested by Blackwelder (1982), who recommended specifying H_O and the alternative H_A so that, "Type I error α[4] and Type II error β are reversed from the case of the usual null hypothesis" (p. 349). Such an arrangement leads to testing the null hypothesis that the standard treatment (conventional training in our case) is more effective than the experimental treatment (new training) by a specified amount, δ. Rejecting H_O, that conventional training is more effective than new training by δ or more, and accepting H_A, that the effects of conventional

[1]Erroneous acceptance of H_O is known as Type II error, which is a central concept throughout much of this book.
[2]See Elementary Rule 2.
[3]See Elementary Rule 13.
[4]Type I error, α, is the risk of erroneously detecting mean differences between the scores of compared groups.

training and new training differ by less than δ, are conclusions with which evaluators are likely to be comfortable. Because α is routinely specified, these test procedures are consistent with traditional hypothesis testing (Blackwelder, 1982).

2. *Specify the risk the evaluation customer is willing to take of erroneously detecting no differences between the compared groups' scores.*

Erroneously finding no statistically significant differences between the scores of compared groups defines Type II error. The probability of Type II error is called β.[5] By specifying β, evaluators, researchers, or test proponents set the risk they are willing to take of making a Type II error. The utility of specifying β can be seen by imagining how our willingness to act on a null result would differ with, say, $\beta = 0.80$ or $\beta = 0.20$: With the lower chance of error at $\beta = .20$, our temptation to base decisions on the null result would be much greater than with $\beta = .80$.

Power = .80 and $\beta = .20$ are not mandatory values any more than is $\alpha = .05$. Policy makers can adjust those values depending on the importance (as defined by costs and hazards) of errors in decisions that will ensue from failing to detect between-group differences and erroneously detecting between-group differences.

Specifying β allows us to accept H_O the same way we routinely accept H_A, that is, with the

[5]As will be elaborated in Elementary Rule 3, $1 - \beta$ is therefore the probability of correctly finding a statistically significant difference and is called the power of the test. With $\beta = .20$, for example, we have 1 - .20 = .80 power, that is, an 80% probability of correctly finding a statistically significant difference between the mean scores of compared groups.

understanding that in approximately 100β times in every 100 tests, for a given effect size, we will be wrong. With β and α specified, evaluators can avoid the error of automatically equating null results with equal effectiveness by stating their conclusion in the following general form: "With $\alpha = X$, $\beta = Y$, and the effect size = Z, we found no statistically significant differences between the compared groups' scores." Presenting null results in other than that general form is in our view a disservice to the evaluation customer and invites suspicion of incomplete analysis at best or chicanery at worst.

We urge evaluators and their customers to consider the relative threats of Type I error and Type II error: Decisions made on erroneous findings of statistically significant superiority of old training over new training (i.e., Type I error) might waste training resources, but seem not to pose much threat to readiness. Erroneous findings of *no* statistically significant differences, however, seem likely to lead to conclusions of "equal effectiveness," with undesired effects on readiness, the national defense, and the vitality of our offspring in armed conflict. Evaluators of military training and their customers should therefore consider setting β at a lower value than the customary 20%, perhaps even as low as 5%.

3. *Perform power analyses to determine the number of observations necessary to detect differences between the scores of compared groups.*

The power of a statistical test is the probability that the test will find an effect, that is, a difference between the mean scores of compared groups in the case of 2-group comparisons, given that an effect of a certain size exists. Without sufficient power, real differences between the proficiency of the compared groups will go undetected. Power is a function of

four quantities: (1) sample size, (2) variance within compared groups' scores, (3) effect size, that is, the size of the actual difference between compared groups' scores, and (4) alpha.[6 7]

By manipulating the last three of these four quantities, we may estimate the first -- that is, the sample size necessary to detect a difference of a given size between the mean scores of compared groups.[8] Doing so is essential in reviewing and evaluating training-evaluation plans, whose adequacy rests in large measure on whether the proposed samples are large enough to allow detection of between-group differences that result from the compared kinds of training. For similar reasons, power analyses are essential in reviewing evaluation reports and assessing the validity of inferences therein -- especially reports in which evaluators infer equal effectiveness of compared kinds of training from null results. The validity of the equal-effectiveness inference depends on whether the null results were due to small or no differences between the scores of compared groups or to sample sizes that had not a prayer of detecting meaningful differences.[9]

---

[6]See Elementary Rule 1 for an introduction to alpha and Shavelson (1988) for considerations in selecting alpha levels.

[7]The four quantities are, in turn, governed by various factors. A summary is in Appendix A.

[8]Computation procedures are in Boldovici and Kolasinski (1997), Cohen (1962), and in references cited therein.

[9]An example is Brown, Pishel, and Southard's (1988) finding of no statistically significant difference between the Army Training and Evaluation Program (ARTEP) scores of four SIMNET-trained platoons and four field-trained platoons. Analyses by Boldovici and Bessemer (1994) revealed statistical power of Brown et al.'s test < .25; that is, this SIMNET evaluation had less than a 25% chance of detecting between-group differences of the size examined. Similarly, Boldovici and Kolasinski (1997) demonstrated that comparisons planned for CCTT with 4 companies in each of two groups would have power ≈ .20 for detecting a 10% difference between the scores of the compared groups, and power ≈ .50 for detecting a 20% difference. That is,

The obvious way to mitigate the embarrassment associated with discovering that results used to tout equal effectiveness of new and old training were due to inadequate sample sizes is to estimate adequate sample sizes by doing power analyses before the evaluation begins.

Before performing a statistical power analysis, an evaluator should have a good numerical estimate of how much the performance measure varies among the sampling units within groups. Within-group (error) variance is one of the quantities needed to compute power, and the lack of this quantity may act as an obstacle deterring the analysis.[10] A determined evaluator will find a way to overcome this obstacle. Results of previous tests using the same or similar measures may provide data yielding a variance estimate. Data collection procedures often can piggyback with little added cost on regular ongoing training exercises that provide an opportunity to obtain similar measures for a sample of units. The evaluator may conduct a pilot test to verify procedures and provide a variance estimate.[11]

Selecting sample sizes that are neither so small as to preclude finding differences between compared groups' scores nor so large as to waste evaluation resources is a straightforward matter whose implementation can save money. If on the one hand, we plan multi-million dollar field trials and the power

---

with $n = 4$ companies per group, the risk of the evaluation's failing to detect a 10% difference between the compared groups' scores was 80%, and the risk of failing to detect a 20% difference between the compared groups' scores was 50%. Even with n tripled to 12 companies per group, the chances of failing to detect a 10% difference between the compared groups' scores was 50%.

[10]For rare cases in which no variability estimates are available or forthcoming, power may be estimated with standard deviation units. Computational details and an example with Army companies are in Boldovici and Kolasinski (1997, p.130).

[11]See Chapter I: Rationalizations.

of the tests is unknown, then we possibly waste the cost of the test. If we compute the power of the test and find it too weak to reveal existing differences between the scores of the compared groups, and we conduct the test as-is, then we waste the cost of the test with certainty. If on the other hand, we find the computed power of a field trial to be in the mid- to high-nineties, the sample size is in the zone of diminishing returns on power; policy makers may then choose to reduce the sample size and save attendant costs.

The costs of scrimping or squandering sample sizes increase as the focus of our training evaluations moves up echelons. Test costs usually grow as the sampling units increase from individual crewmen or commander-gunner pairs in tank-gunnery trainers, through crews and platoons for Simulation Networking (SIMNET), to companies and eventually battalions for the CCTT. The costs grow because sample-size requirements remain similar regardless of which echelon we use as the test unit. Unless the variation in performance measures or the desired effect sizes differ greatly among echelons, tests must sample about the same number of elements at each level to obtain adequate statistical power. Although a battalion has more personnel than a platoon, each battalion is only one sampling unit if the test measures performance at the battalion level. Conducting individual- or crew-level tests with insufficient power to detect differences between groups' scores may be rationalized as a negligible waste of evaluation resources. Conducting similar tests to compare groups of companies or battalions is an unconscionable waste of evaluation resources.

The way to avoid costly errors such as hypothesized above is to do power analyses before comparing the effects of conventional and new training. Results of the power analyses will tell us, with given sample sizes, the probability of finding differences that exist

between the scores of compared groups. With knowledge of the capability of our evaluation to find real differences between the compared groups' scores, we can make informed decisions about whether to spend the money required to conduct the evaluations. Consider, for example, how our decision about whether to conduct a comparison between new training and conventional training might differ depending on whether the power analyses told us we had a 5% chance or a 95% chance of detecting real differences between compared groups' scores. Such informed decisions have, to the best of our knowledge, never been made in planning evaluations for new Army training; the power analyses were not done.[12]

4. *Increase power by reducing the variability of performance measures within the compared groups.*

When the computed power for a prospective training evaluation test turns out to be inadequate even with the maximum possible sample size that can be obtained, we can only increase power by making the variance smaller. Before finally deciding to abandon the test, evaluators should therefore examine some possible changes in test design and procedures to reduce variance. This method of increasing power is more difficult to manage than simply increasing sample size. Effective variance reduction requires understanding of many sources of variation in performance, as well as knowledge of feasible techniques that reduce or eliminate the influence of various sources.

Four sources may contribute to within-group variation in performance measures: (a) treatment

[12]As is the case with proving H_0, we realize the impossibility of proving that no power analyses were done and welcome evidence to the contrary.

variations, (b) environmental variations, (c) sampling-unit characteristics, and (d) measurement reliability.

(a) Treatment variations are, in the case of training, differences among units in the administration of training events or procedures. Evaluators sometimes have allowed units to select exercises they prefer or to create their own exercises, so that each unit teaches somewhat different tasks or similar tasks under different conditions. Even when the exercises are the same for all units, trainers may deviate from procedures specified in their training support packages. Allotted time may run out, cutting short part of the training, for example. And trainers may differ in techniques and skills in conducting after action reviews.

(b) Environmental variations are changes in features of the setting where units conduct training or concurrent events outside the training context that may affect some units and not others. Equipment failures, either in field or simulator contexts, may affect some units' training more than others. Weather conditions are an important factor affecting visibility or maneuverability of units in the field. High-ranking visitors to new simulator facilities may increase pressures to perform for units training at the time. Units may be required to assign personnel to work details or other duties that force them to miss some or all of the training.

(c) Units sampled for membership in the treatment groups differ in numerous ways that may affect performance measures. The knowledge, skills, and abilities of individual unit personnel influence unit performance, especially those individuals holding positions of leadership. Many unit-level characteristics, such as

cohesion, command climate, personnel turbulence, training history, and shared experiences may differ substantially among units and contribute to performance variations.

(d) Reliability of measures refers to the consistency of scores obtained for the same performance measure on different occasions, or agreement among parts of a composite measure. Variations that contribute to unreliability may come from the measuring instrument, procedures, conditions, or unit performance. In Army exercises, observers often serve as measuring instruments. Observers may differ in their standards for judging performance, and the same observer may be inconsistent on different occasions. Observers may change systematically as they gain experience with multiple units across the duration of a field trial. Training and experience may make observer standards and procedures more similar and each observer more consistent. Observers may view an exercise from different vantage points or with different sensors, thus sampling different aspects of performance. Visibility conditions may affect observation, and units' performance levels simply may change from one occasion to another. Unit proficiency may differ between various performance elements, producing inconsistency among parts of a composite measure.

To the extent we reduce the within-group variance from any of the four sources noted above, statistical power will increase.[13]

Cook and Campbell (1979) presented a number of specific methods for reducing within-group variance (p. 47-49). In essence, these methods use four

[13]Also see Elementary Rule 9.

techniques to curb variables' influence on the performance measure: (a) hold constant, (b) deliberately vary, (c) measure and equate, or (d) measure and adjust. Appendix B is a summary of these four techniques.

5. *Randomly assign soldiers or units to the compared kinds of training.*

Randomization insures that each individual or unit has equal probability of assignment to the treatment groups. This creates "statistical equality" between the effects of all sampling unit characteristics assigned to both groups. Statistically, the expected mean averaged over all possible samples of such effects is the same for both groups. The treatment comparison is therefore unaffected except by the treatment effect, kinds of training in our case, if such an effect exists.

Consider, for example, a case in which evaluators have completed a field trial, analyzed the data, and found a statistically significant difference between the two kinds of compared training, for example, new training and conventional training. What conclusion can we draw from this result? We planned and carried out the project to compare the kinds of training, so the obvious conclusion is that the kinds of training ("treatments") caused the difference. More specifically, evaluators may infer that one treatment produced better learning and transfer than the other treatment. Attributing the cause of the observed effect to the treatments alone is, however, a valid inference only if no alternative cause or causes can explain the result. Suppose evaluators in this example assigned intact units to the new and old training. Reasons for doing so might include administrative convenience or overcoming commanders' objections that random assignment violates unit integrity. All platoons in one battalion therefore got assigned to one

treatment group, and all platoons in another battalion got assigned to the other treatment group. This procedure insures that all unit characteristics shared by platoons in a battalion but that differ between battalions, for example, platoon SOPs or recent training history, will be confounded with treatments, and the results will be biased.

6. *Randomize variables whose effects cannot be controlled or measured.*

Suppose our compared groups differ by one or more variables in addition to the kinds of training they receive, for example, one of the compared groups receives more training than the other. If these other variables affect performance, then any observed difference between the compared groups' performance confounds treatment effects with the effects of the other variables – the effect of amount of training, for example. Such confounding of effects results in a biased and therefore inaccurate estimate of the treatment effect. In fact, the other variables could be entirely responsible for the observed difference, while the treatments actually contribute nothing to the difference. To insure valid inference about the treatment effect, evaluators must prevent the occurrence of confounded treatments and attendant biased results.

The risk of confounding treatment effects also is diminished by arranging the sampling units for training and testing in a random order. This procedure randomly associates any effects of environmental and measurement conditions with treatment groups, thus equating these effects statistically between or among groups, and thereby diminishing confounding and bias.

Despite the desirability of randomization, practical or political considerations may preclude its use or allow only partial use in many circumstances. In the

face of known risks to valid inference, responsible Army authorities often will feel compelled to proceed with some kind of evaluation as justification for prior decisions and expenditures. These situations force evaluators to turn to backup methods in an attempt to salvage something of value in return for the resources expended in a field trial. With judicious application, several alternatives may be considered that are usable under special conditions, allow partially valid inferences subject to specific caveats, or allow valid inferences for limited questions while postponing decisions that require broader issues to be addressed. We discuss a number of these alternative methods in Chapter IV: Suggestions.

7. *Establish that the compared groups do not differ in ways that might affect outcomes and alter conclusions.*

The reason for not wanting pre-existing differences between groups is, of course, that the pre-existing differences may influence the outcomes of our evaluations and will make figuring out what caused the evaluation results difficult and perhaps impossible. Suppose one of the compared groups comprises units that have had more Combat Training Center rotations than the other group, for example, and the greater CTC-rotation group scores significantly better in a transfer evaluation than do the units in the lower-rotation group. We would be hard-pressed to ascertain whether the difference between the compared groups' transfer scores was due to the different kinds of training the groups received during our evaluation or due to the CTC-rotation differences that existed between the groups before our evaluation began.

As discussed in previous rules, a good way to diminish pre-evaluation differences between compared groups is to assign soldiers or units randomly to the groups. Nevertheless, random

assignment may not be possible when the compared groups are Army units; field trials of Army training often compare scores of intact units. Drawing any valid inferences from outcomes based on intact units requires additional information to determine how the units differ between groups, and to estimate how these differences may have affected evaluation outcomes.

Even with random assignment of units to treatments in training-effectiveness evaluations, some pre-evaluation proficiency differences between groups will still exist resulting from variations among units assigned to the groups. It makes little sense to rely entirely on the beneficial effects of random assignment when we can examine those effects empirically. One way to test our assumption about the compared groups' equality is by the use of pretests (pre-training tests). The pretest can be the same as the transfer test (post-training test or posttest). The pretest also may differ from the transfer test, but the performance measured by the pretest must have a strong relationship to the performance measured by the posttest.

Analyses of scores on a pretest will tell us whether the compared groups' proficiencies do or do not differ significantly before our evaluation begins. If we find no significant pre-evaluation proficiency differences between the compared groups coupled with a narrow confidence interval,[14] then we suspect random assignment of units to groups had its intended effect. If we do find significant pre-evaluation differences among the groups, then our choices are either to use a covariance analysis or to deliberately equate the groups by matching units on pretest performance.[15]

---

[14]See Elementary Rule 15 re. confidence intervals.
[15]Details are in Section IV: Suggestions.

The downside of using pretests is that members of the compared groups learn something as the result of taking the pretests. When such learning occurs, we cannot separate gains in posttest scores resulting from the groups' practice during the pretest from gains resulting from training administered during our evaluation. The way out of that predicament is to use an additional group that takes the pretests and posttests but receives no training, that is, a no-training control group.[16] Data from the no-training control group permit separating the amount of the compared groups' proficiency acquired during pretesting from the amount of proficiency due to training.

A pretest also may remind group members of forgotten knowledge useful in training, or it may sensitize them to facets of the training conditions that then changes their reaction to, and the results of, training. To examine such pretest-treatment interaction effects, the evaluation design must be expanded to include treatment and control groups that are not pretested. The analysis and interpretation of such designs is as for the Solomon four-group design described by Campbell and Stanley (1963).

One way to avoid using pretests and the attendant burden of a no-training control group is to match units in the compared groups based on a variable other than pretest scores. Examples of such variables include mean numbers of CTC rotations, expert ratings, or other proficiency estimates. The object in such an assignment method is, of course, to minimize between-group differences on the

[16]Safety considerations may preclude use of no-training control groups with novices. For sustainment training that is often the purpose of simulator training, however, a no-training control group will usually be feasible. To avoid leaving the control group at a permanent disadvantage, the missed training can be given after the posttest.

variables of interest before the evaluation begins. We discuss this kind of matching in Chapter IV: Suggestions.

8. *Equalize or systematically vary the amount of training provided by treatments to prevent confounding with the treatment comparison.*

As noted earlier, when the compared groups in a training evaluation are treated differently in ways other than the training method, then the treatments, kinds of training, are said to be confounded. Confounded treatments make it difficult or impossible to determine what caused observed differences in performance among treatment groups. Military training evaluations, especially those that compare device-based with conventional training, often confound treatments with amounts of training. In training evaluations with aircraft simulators, for example, the compared groups usually are trained to some criterion of proficiency in simulators before being transferred to the aircraft.[17] If the compared groups take different amounts of time, trials, or both to reach the criterion of proficiency in the simulator, then the treatment, that is, kind of training, is confounded by amount of training. The confounding may not be of concern to aviation trainers and evaluators inasmuch as the only choice available is between the simulator and the aircraft, and one intended benefit of the simulator is increased opportunity for practice. If the simulator training proves equal or superior to conventional training,

[17]Safety considerations notwithstanding, the rationale for training to criterion in the simulator seems grounded in evaluators' desire to make the aviators' proficiency equal before administering trials in the aircraft; this creates the illusion of fairness in the test. In reality, insuring equal proficiency in such cases is impossible: When all aviators reach criterion, their scores are the same because proficiency above criterion is not measured. All individuals end their practice performing at criterion level, but with an unknown distribution of proficiencies above criterion.

then it matters not whether the training medium, simulation in this case, caused this effect, or the cause was an increased amount of training.

In some cases of Army training, such as tactical and maneuver training for example, the consequences of confounding kinds and amounts of training are different than they are in aviation. Suppose compared groups with different training treatments also receive different amounts of practice. If the resulting proficiency of the group receiving more practice exceeds that of the group receiving less, then the results cannot legitimately be attributed to the compared *kinds* of training. The results are more parsimoniously interpreted in terms of *amounts* of training. That is, more training was simply better than less. Such a result, in the case of simulator-based Army tactical training for example, tells nothing about the efficacy of the new simulation and therefore fails to support a purchase decision. Similar results could have ensued from using increased amounts of training with, for example, classroom instruction, sand tables, or other simulations less expensive than the simulation under examination.[18]

The obvious way to avoid confounding kinds of training with amounts of training is to give all subjects in all compared groups identical numbers of trials or identical amounts of practice time. The evaluator must choose to control trials or time, because holding both constant is impossible. If the number of trials is fixed, the time required to complete the trials will necessarily differ among sampling units. And if training time is fixed, the number of trials that can be completed in that time will necessarily differ among sampling units. The evaluator must base the choice on the appropriate

[18]To preclude this argument, evaluators should include some of these cheap alternative training methods for comparison.

measure of practice for the kind of task or skill defined as the training objective. For tasks performed continuously at every moment of the practice session such as vehicle driving, time is the measure of amount of practice. For tasks performed with discrete actions, such as procedures or decisions, trials (i.e., task repetitions) are the measure of amount of practice.

Holding the amount of training constant has a downside. The treatment difference observed in a subsequent performance test may depend on the specific amount of training provided during our evaluation. This limits the generality of inferences possible about the relative effectiveness of the compared kinds of training and requires evaluators to qualify their conclusions. The only way to avoid this limitation is by systematically varying the amount of training and determining empirically how the treatment effects change as a function of amount of training.

9. *Allow some time to pass after the end of training before giving posttests, and equalize this time for the compared groups.*

Aside from administrative convenience, one deceptively attractive motive to administer transfer tests immediately after training is a belief that doing so will reveal the maximum amount of transfer attributable to training. A common expectation is that an immediate posttest will yield a high-water benchmark that later transfer scores are unlikely to exceed. That line of thinking is erroneous because the measured amount of transfer and even its direction often change over time and with intervening experience. Performance on early transfer trials is not necessarily a good predictor of performance on later transfer trials. Testing immediately after training often will yield scores quite different from the results of testing later. The

amount of transfer will therefore change as time increases between training and transfer testing.

The amount of transfer may improve on delayed transfer tests because individuals continue to study task procedures and ruminate about their mistakes after training, or group members discuss ways to improve their units' performance. Positive transfer may decline because forgetting what was learned in training reduces transfer. Such changes can depend on the amount of training and original learning with new training. Greater amounts of training and learning tend to produce transfer that is more durable. Forgetting can even change negative transfer to positive transfer. This can happen when aspects of performance that produced negative transfer shortly after training are forgotten, leaving intact other aspects of performance that contribute to positive transfer. One-shot, two-group, combined-arms, multi-echelon evaluations typically do not manipulate the amount of new training, the amount of training or testing in the transfer situation, or the intervening time. Because all results depend on the specific values of these variables, the generality of all results is open to question.

10. Use only performance tests whose scores are reliable.

Reliability refers to the dependability or repeatability of test scores that distinguish between superior and inferior performances (Gagné, 1954). As noted in Chapter I, reliability coefficients range from 0.0, which indicates total inconsistency in the ability of scores to discriminate, to 1.00, which indicates total consistency.

Standards for test reliability depend on how the test is to be used and other circumstances associated with that use. The reliability of tests used to evaluate training with large samples can be as low as .50.

With moderate or small samples, test reliability should be .70 or greater. If test scores feed decisions about particular individuals or units, reliability should be .90 or greater.

One reason reliability is important is its effect on the statistical power of tests: As noted in Elementary Rule 3, power is the probability of detecting true differences between the scores of compared groups. Power decreases with decreased measurement reliability, because of increased error variance. Classical test theory shows that increasing test length is an easy way to increase reliability, when the additional items correlate well with the original test items. Doubling a test with reliability .50 will, for example, increase the reliability to .68.[19]

Performance scores that do not meet standards of reliability and validity do not qualify as indicators of training effectiveness. In advance of any field trial of new training, evaluators should therefore conduct a tryout of the tests to insure that the measurement and data collection procedures are workable. Data

[19] If the reliability of the original test is r_{11}, and the number of test items increases by a factor of k, then the reliability of the lengthened test is $r_{kk} = kr_{11} / 1 + (k - 1) r_{11}$. In an evaluation design with only a posttest after training, doubling our example test with .50 reliability has the effect of reducing the within-group error variance (σ_y^2) by one-sixth. Power also may be increased in a covariance analysis of a pretest-posttest design by lengthening the posttest more than the pretest. With the total number of pretest and posttest items held constant, the optimum allocation of items will make the posttest longer by a factor of K, while making the pretest shorter by a factor of $2 - K$. The following formula determines the value factor K: $k = (1 + r_{yy}) / (r_{xy} + r_{yy})$. The pretest-posttest correlation (r_{xy}) and the reliability (r_{yy}) must be estimated for equal-length tests in a pilot trial. If a pilot trial is not possible, the evaluator can estimate approximations to these values by using available data for similar tests that are regularly used to evaluate training at the test site. For example, with r_{xy} and r_{yy} both equal .50, then $k = 1.5$ and $2 - K = .50$, and the posttest should have three times as many items as the pretest. As a rule of thumb, Maxwell (1994) suggested that this ratio often is near optimal in most behavioral research.

obtained in the tryout provide a base for examining reliability and validity. If a tryout is not possible, evaluators can obtain surrogate estimates based on examining available data from similar tests used in regularly ongoing training. Advance verification is required to insure that the quality of the tests is adequate for the purposes of the trial. To do otherwise runs a risk of wasting resources on a trial with no useful results, or worse, with results that mislead.

One way to estimate reliability is by computing an internal consistency coefficient based on the average intercorrelation among the test items (Nunnally, 1967). A number of formulas exist for this estimate depending on the nature of the test items and the scoring procedures.[20] If ratings are used, computation of inter-rater reliability coefficients is appropriate.[21] Another way to estimate test reliability is to correlate scores obtained from two or more administrations of the test to the same subjects. In addition to measurement errors, changes that occur between the administrations affect such test-retest reliability coefficients. A test-retest estimate is therefore less pure as a measure of reliability than an internal consistency coefficient. If the time interval is long, for example, forgetting will cause performance changes. Retesting is usually not possible in large-scale, combined-arms, multi-echelon field trials of unit training. As will be seen later, however, obtaining repeated measures of the compared groups' transfer performance is valuable for reasons beyond estimating measurement reliability.

---

[20]For cases in which only one set of scores or ratings is available, we suggest consulting an expert in methods of psychological testing and educational evaluation to determine how to estimate reliability and validity for the performance measures used in the evaluation.

[21]The use of scores from ratings is discussed in Chapter III. Methods for estimating reliability and validity are in Appendix C.

Reliability increases with removal of test items that correlate poorly with other items in the test. Because internal consistency reliability depends on the size of item intercorrelations, a homogeneous test composed of similar items that tap into the same underlying skill will tend to maximize reliability. Deleting items usually will not be a desirable option for many Army performance tests, because procedural, tactical, or decision-making tasks often involve complex combinations of different kinds of task elements requiring many skills. Instead of removing items, the better course may be to sort different groups of items into subtests composed of similar items that correlate well within the group. Then each subtest may have high reliability, and the multiple subtests retain the coverage of task elements required for combat proficiency. Groups of task elements in unit tactical exercises, for example, might form subtests relating to communication, maneuver, engagement, and sustainment. Such subtests are more instructive about the outcomes of training than is a total score based on all items. Results might, for example, show simulator training to be more beneficial for communications and maneuver performance than for the other task elements.

11. Report test reliabilities and their implications for validity.

Validity of a test involves how well it represents the property it is intended to measure and fulfills the purpose for which that measure is obtained. Validity has several definitions relating to different purposes, each with different methods for estimating validity.[22] For performance tests used in field trials to evaluate training, the main

[22]See, for example, Wilkinson, L. and Task Force on Scientific Inference, APA Board of Scientific Affairs (1999).

requirements are for both content and predictive validity. Content validity concerns how well the test represents the domain of performance included in the training. This kind of validity is easy to demonstrate for Army performance tests in advance of their use in a field trial. Well documented task analyses performed by subject-matter experts for the tasks included in the training and in the test are available in Army publications to establish content validity. The tasks represent the desired performance domain if they are included in the Mission Essential Task Lists (METL) for the kind of unit that forms the target population for the training. The final condition required for content validity is that the measurement procedure successfully captures the performance of the selected tasks or task elements.

Predictive validity refers to the ability of test scores to relate to and forecast later measures of important behavior. The later measure is often termed the validation criterion. Measures obtained in subsequent training or exercise events provide criteria for partially validating Army performance tests. If, for example, a performance measure obtained in a simulator test can predict performance in a field exercise, this provides evidence of test validity for that criterion measure. Similarly, validity of a field test measure can be assessed by relating it to a criterion measure from a Combat Training Center exercise. The ultimate criterion of validity for tactical tests is of course combat performance, and measures of this kind are virtually unobtainable.

Aside from the effect of reliability on power, reliability is important because of its relation to predictive validity. On the one hand, a test may be highly reliable, yet have little utility because it does not measure the intended dimension of behavior or have predictive validity. Two raters, for example, might show 100% agreement between their assigned

ratings, and both raters could be wrong about the quality of performance. On the other hand, reliability is important because it sets a limit on predictive validity both logically and statistically. To the extent that scores are composed of random errors of measurement, a boundary is set on the portion of scores available for valid performance measurement. As Ghiselli (1964) showed, a correlation coefficient measuring the degree of predictive validity for a test cannot exceed the square root of the reliability coefficient for that test.[23] High reliability does not guarantee high validity. But without some reliability, inferences about estimates of transfer of training from test results cannot be valid.[24]

12. *Avoid ceiling and floor effects by adjusting posttest difficulty to produce scores between 75% to 25%, or use more than one posttest with varied levels of difficulty.*

Evaluations that use tests bounded at high and low ends may be vulnerable to ceiling and floor effects. Percentage scales used to assess unit performance, such as percentage of tasks correctly performed, have bounds at 0% and 100% and are subject to these effects. Ceiling and floor effects create limitations on the relations between the observed scores and the ability levels that might be inferred from the scores. A 5% difference between unit scores of 50% and 55% may represent a small ability difference, but a difference between 90% and 95% may indicate a much larger ability difference.

Ceiling effects happen when the compared groups score high in performing the transfer tasks. All scores fall near 100% with little room for the test to

[23]See Appendix C for elaboration of this point.
[24]More exactly, an inference can be no more valid than would be expected by chance.

discriminate among compared groups. Finding any statistically reliable differences between the compared groups' transfer scores thus becomes extremely unlikely even if such differences exist. Ceiling effects can happen because training was extremely effective, because all subjects were proficient on the transfer tasks before training began, or for any other reason that makes performing the transfer tasks easy. Scores at or near the maximum are, of course, desirable if you are a trainer. Scores at or near the maximum are not, for the reason mentioned above, desirable if you are an evaluator.

In contrast to ceiling effects, floor effects happen when the compared groups score low in performing the transfer tasks. In most training evaluations, very low scores and floor effects are likely only on pretests with units previously untrained, when tasks are novel and unlike other tasks previously learned. Floor effects are possible on posttests only when the amount of training has been grossly inadequate or the training is entirely ineffective – a combination of conditions unlikely to occur in practice. If floor effects do occur, however, they will mask proficiency differences between the compared groups, either of which may have been superior if they had had more training.

Discrimination among group scores is best when the difficulty of items is around 50%. Training evaluators should aim for mean transfer scores around 50%, rarely higher than 75% or lower than 25%. Otherwise variability is likely to be restricted to the point where finding significant differences between the compared groups' scores is impossible.

Implications for using experienced units in training evaluations seem worth considering. How do we design evaluations so that experienced units score at or near 50% on a transfer test? Using recent AIT

graduates for CCTT evaluations might work where safety considerations are not a problem. But doing so would invite the charge that our subject sample was not representative of the population of interest. Resolving this issue depends on which evaluation question we wish to answer: Are we interested in determining whether using the new training vs. the old has differential effects for any population at all? If the answer is no, then the use of tyros is contraindicated. Or are we interested in determining whether using the new training vs. old for sustainment has differential effects with experienced units? If the answer is yes, then special tests must be designed, tried out, and revised to adjust the difficulty of tasks until the necessary 25% to 75% distribution of scores is achieved. Every effort must be taken to avoid ceiling effects, because they mask proficiency differences between compared groups, either of which may have been superior had the transfer test been more difficult.

Where practical constraints, such as lack of time or resources, preclude posttest development in advance, the only alternative is to use two or more posttests with deliberate manipulation of task difficulties. A baseline standard test can contain tasks with normal task conditions and standards. One or more variants can be created with conditions and standards adjusted to increase difficulty. With tactical tasks, for example, conditions such as visibility (smoke, night), enemy (numbers, weapons, ability), terrain (maneuver space, obstacles, cover, lines of sight, vegetation, buildings), and communication (jamming, terrain) can be adjusted to increase task difficulty. This approach has the benefit of providing an examination of the resiliency of transfer performance under difficult conditions likely to occur in combat. If transfer from new training remains evident under difficult conditions, even if it is diminished in size, our confidence in the value of that training for combat improves.

Using two or more posttests requires using experimental designs that are complicated. Different samples of units in each treatment condition may get different posttests, requiring a larger total sample size. Alternatively, each unit may get all tests, but given in different orders.[25] In either case, the use of more than one posttest provides repeated measures of proficiency that can have a beneficial side effect by increasing statistical power.

Another method could be valuable in avoiding ceiling effects. Item Response Theory (IRT) is an alternative to classical test theory that relates probabilities of response to test items to a theoretical scale of ability of performers and difficulty of test items (Lord, 1980; Embretson & Hershbeger, 1999). Analysis of test data using IRT models estimates the position of both test takers and test items on a common linear scale. Measures derived by IRT models have the powerful property of making performers' scores independent of item difficulty, satisfying the principle of additive conjoint measurement required for basic measurement of human performance (Luce & Tukey, 1964). Scores on the linear scale are unbounded and therefore free of ceiling effects. The main obstacle to using IRT models is that large samples of individuals or units (200 or more in most cases) are required to produce stable estimates. IRT methods become usable for Army training evaluations only with performance data accumulated and archived over a long period of time in a stable training environment where the same tests are used over and over.

[25]See discussion of Latin Squares in Chapter IV and Appendix D.

13. Use conventional analyses of raw scores to estimate training effects.

Various analyses of data from transfer evaluations lead to spurious inferences. Correlations between training scores and test scores, for example, yield only weak inferences, because correlations do not establish the causal link necessary for demonstrating transfer. A high positive correlation between training scores and test scores suggests only that subjects used similar skills in training and in testing. A high positive correlation between training scores and test scores does not demonstrate that training caused the test scores.

The results of transfer formulas and transfer-efficiency or savings measures also are misleading. The reasons for this relate to deficiencies that inhere in difference scores, percentages, savings estimates, and unit pricing. Details are in Appendix E.

Avoid using transfer formulas, correlation, efficiency and savings for estimating transfer. If for some reason you must use them, be sure to supplement them with conventional analyses of raw scores. Conventional analyses include *t*-tests, analyses of variance, confidence intervals, and simple-effects tests if the number of compared groups is three or more. Failure to use conventional analyses of raw scores will produce misleading results for reasons elaborated in Appendix E.

14. Perform separate analyses of training-sensitive and training-insensitive test items.

Some tasks performed by Army units are insensitive to new, device-based training; such tasks include stringing wires among stationary tanks for inter-crew commo, road marches, and latrine digging. Unbiased interpretation of evaluation results

requires that analyses be done separately for training-sensitive and training-insensitive tasks. The common practice of averaging scores for both kinds of tasks masks differential effects resulting from the compared training regimens, biases the comparison in favor of null results, and sets the stage for the unwarranted inductive leap to "equally effective" training. To see why this is so, consider the nearly universal case in which field training is better than device training for some tasks and worse for others. A comparison of mean scores for all tasks between field training and device training would tend to wash out both the strengths and the weakness of device training and of field training: If the proportions of tasks favoring field training and tasks favoring device training are similar, the average scores for the two media will show small differences; if the proportions are identical, the average scores for the two kinds of training will be identical.

15. Interpret null results in terms of confidence intervals and power analyses.

If a training evaluation yields no statistically significant differences between compared groups' scores, then the evaluators should calculate confidence intervals to estimate the likelihood that the null results were due to the absence of a proficiency difference between the groups or to evaluation deficiencies that precluded detecting differences. That is, when any test of H_0, including training-evaluation tests of H_0,[26] yields null results, a conclusion of equal effectiveness never follows automatically.[27] As mentioned in Elementary Rule 1 and belabored throughout this book, equating null results with equal effectiveness is an unwarranted inductive leap. An easy way to examine the tenability

[26]See various SIMNET and CCTT evaluations for examples.
[27]EquivTest software to support establishing equal effectiveness of treatments is at http://www.statsolusa.com/

of an equal-effectiveness conclusion is by calculating and interpreting confidence intervals.[28] Without supporting evidence from confidence intervals, a conclusion of equal effectiveness applied to new and old training is untenable.

Confidence intervals differ from hypothesis tests but are closely related. If the same α is used, then the decision to reject or not to reject H_0 will be the same whether a confidence interval or a hypothesis test is used. The advantage of using confidence intervals is that, in addition to permitting hypothesis testing, confidence intervals bound the observed difference between compared groups' mean scores: "A hypothesis test tells us whether the observed data are consistent with the null hypothesis, and a confidence interval tells us which hypotheses are consistent with the data" (Blackwelder, 1982, p. 350). That is, the confidence interval displays the set of differences that are plausible given the data obtained. For a 100 $(1 - \alpha)$ % confidence interval, the conclusion is, "We can be 100 $(1 - \alpha)$ % confident that the interval contains the true value of the difference between the compared groups' mean scores." A wide confidence interval, as compared to a narrow confidence interval, indicates that a greater proportion of the range of differences between the compared groups' mean scores is included in the interval. A wide interval contraindicates equal effectiveness of compared training regimens, even when the mean transfer scores of the compared groups show no statistically significant differences. A narrow confidence interval on the other hand indicates fewer possible values for the difference between compared groups' means than does a wide

[28]We know of no Army training evaluation in which null results were interpreted in terms of confidence intervals. As is the case with proving H_0, we realize the impossibility of proving no confidence intervals ever were reported and welcome evidence to the contrary.

interval and suggests the possibility of – but never proves – equal effectiveness.[29]

In addition to reporting confidence intervals, evaluators who find no transfer differences between compared groups should report the results of power analyses. As noted in Elementary Rule 3, a power analysis done before an evaluation will estimate the probability that the evaluation is capable of detecting group differences of a specified size. The benefit of discovering inadequate statistical power before the evaluation begins is obvious: We can change the evaluation design to obtain adequate statistical power, or we can abort to avoid wasting money.

If pre-evaluation analyses suggest adequate test power but analyses of our evaluation results reveal no statistically significant differences due to the compared training alternatives, we recommend doing power analyses using data collected during the evaluation.[30] Reporting the results of power

---

[29]Boldovici and Kolasinski (1997) computed the confidence interval using results of Brown et al.'s (1988) SIMNET evaluation, which found no statistically significant differences between the proficiency of platoons trained conventionally and platoons trained with SIMNET. The confidence interval contained zero, thus supporting Brown et al.'s finding no statistically significant differences between the compared groups' scores. The confidence interval was wide, however: It contained over half the possible differences between the compared groups' scores. We therefore concluded, "Both the hypothesis test and the confidence interval led us not to reject the possibility of equal effectiveness. But the confidence interval provided additional information suggesting a high degree of uncertainty associated with an equal-effectiveness interpretation" (Boldovici & Kolasinski, 1997, p. 133). Our guess is that computing confidence intervals for many, if not all, military training evaluations in which alternative training regimens were declared equally effective [see Orlansky (1985) for examples] would demonstrate those declarations to be wrong.

[30]This view is equivocal. Some statisticians argue that if we find no statistically significant differences between the scores of compared groups, the null result constitutes sufficient evidence

analyses will help readers decide whether statistically nonsignificant differences resulted from small or no differences between the effects of compared training alternatives or from an evaluation design or analysis that was incapable of detecting differences. Performing post-evaluation power analyses also provides a check on the variability estimates used in pre-evaluation power analyses and could lead to accumulating lessons learned for use in designing later evaluations.

16. Address the generality of evaluation results and of attendant inferences.

The reason for wanting to know about the generality of training-evaluation results and the validity of inferences that attend those results derives from the fact that most experiments and all military training evaluations with which we are familiar yield results that constitute a point estimate. That estimate is one of a theoretically infinite number of point estimates that would aggregate from an infinite number of replications of our evaluation to form a distribution of all possible results. We should therefore like to know the extent, if any, to which our results and inferences can be expected to apply to the populations from which the samples we tested were drawn. Without evidence to support the generality of our result, it is gratuitous to assume that the results of any training evaluation bear some systematic relation, that is, have generality for, the population from which our test samples were drawn -- or for that matter that our result has generality for our test

---

we underestimated the power of our test, and pursuing the matter further – as with post-evaluation power analyses – is pointless. Our view is that without post-evaluation power analyses, we shall never know how much our pre-evaluation power analyses were in error, and shall therefore be ill-positioned to ascertain *why* our pre-evaluation power analyses were in error, and thus likely to make the same mistake again.

samples themselves.[31] Generality is an empirical matter. It does not follow automatically from unsupportable beliefs in, for example, the representativeness of our samples. Unsupported belief in the generality of training-evaluation results from test samples to other samples, from test samples to the population of interest, or from test samples to the aggregated theoretically infinite number of point estimates mentioned above is a common inferential error. The question immediately arises therefore, "What evidence shall we require to bolster our confidence in the generality of an evaluation result?" We have no easy answer to this question; we hope that is because there is none.

Statistical tests of significance provide some evidence, albeit indirect, of the generality of a result. If our training evaluation yields mean differences between the scores of compared groups, and the difference is statistically significant at, say, the .001 level, we are confident that this difference is real and is not the result of error variance in the scores that led to the result. Our confidence here is grounded, partly at least, in the belief that our result has a high probability, 99.99% in the present case, of replication.

Replication is central to estimating the generality of results as replicability is the only way anyone will ever know anything about the population to which the evaluator is generalizing.[32] As for the feasibility of replication for large-scale military evaluations of new training, we can only say we have never seen one and think it unlikely we shall. Absent replication, and the weak evidence provided by significance

[31]For elaboration of these points we refer readers to the work of Tversky and Kahnemann (1971), summarized partially in our Chapter 1.
[32]See Thompson (1997) for elaboration.

testing notwithstanding, what approaches remain for addressing the generality of a[illegible]luation result?

One approach to addressing generality is to report the results of reliability estimations and their implications for validity.[33] Split-half reliability, in which half the scores from a test are correlated with the other half, is a poor man's version of replication. Cross validation (Efron, 1982) serves similarly.

Another possible approach to estimating the generality of evaluation results is via resampling methods, which comprise exact permutation tests, bootstrapping, jackknifing, and cross validation.[34] The resampling methods are computer-intensive but are getting increased attention with increased computing power on desktops. Whether and the extent to which these methods are appropriate for estimating the generality of training evaluation results is not clear to us. Because the methods may be applicable in ways we do not know, we suggest that evaluation planners discuss them with authorities on their use and implications.

If, in the final analysis, our results and our analyses permit no tenable inferences about generality, our evaluation plans, reports, and briefings should say so. Candor and commonality of aims between evaluators and members of training-acquisition communities will, we hope, lead to productive discussions of the effects of unknown generality for

[33]See Elementary Rule 11

[34]Free bootstrap software add-on packages for the R statistics program (a GNU licensed variant of S) can be obtained from Comprehensive R Archive Network (CRAN) sites such as http://lib.stat.cmu.edu/R/CRAN/. Bootstrap methods are included in the free Dataplot statistical software available from the National Institute of Standards and Technology (NIST) at http://www.itl.nist.gov/div898/software/dataplot/homepage.htm. Inexpensive and easy-to-use resampling shareware also may be downloaded from http://www.resample.com.

training-acquisition decisions, readiness, and the national defense.

17. Never accept evaluation plans or results at face value.

Our hope is that persons responsible for approving plans for new-training evaluations and that persons responsible for decisions made in light of training-evaluation results will use the rules and other materials in this book as checklists for reviewing evaluation plans and for making inferences from evaluation results. We hope also that Table S-1 will facilitate doing so.

18. Monitor indicators of the value of new training during fielding and implementation over the long term to insure sustained training effectiveness.

No matter how well trained, the trainers and equipment operators who deliver new training with simulators cannot produce results in an initial operational test that are typical of normal long-term training results.[35] As the training is fielded and implemented at various sites, the trainers and operators gain experience, develop habits of action and routine procedures, and institutionalize the training as part of the normal training cycle for local units. The training results may get better as a deeper

[35]Bessemer's (1991) research with SIMNET illustrates the importance of following up on the results of one-time transfer tests administered immediately after training. As instructors gained experience with SIMNET, they became more proficient in using SIMNET for tactical leader training. Transfer from SIMNET training to platoon leader performance in field exercises began emerging three months and five classes after the Armor Officer Basic Course classes first performed platoon-level exercises in SIMNET. Transfer continued to increase gradually in the subsequent seven classes observed over an additional five months. This research suggests that early transfer tests pairing new devices with inexperienced instructors are likely to yield lower transfer scores than would be obtained on later tests.

understanding of how to get the most from new training capabilities develops, or the results may get worse as key points of initial "train the trainer" classes are forgotten. Following some standard indicators of training results may suggest [illegible] the expected benefits of the training are sustained over time. Such results may show need for corrective actions to improve training and to protect the Army's return on investment. The indicators may derive from in-device tests or from standard field test exercises at local training areas. Surveys of user satisfaction and other indicators may prove useful in addition to performance results (Bessemer & Myers, 1998). Besides examining long-term trends, the indicators enable investigation of other questions, such as whether the quality of training is exceptionally good or bad at various sites. Such results also establish a baseline for evaluating the effects of training improvements.

Conclusions and Recommendation

As was shown in Table S-1, the consequences of ignoring any of the elementary rules of design and analysis discussed above include that evaluations (a) may be biased to favor finding no differences among the compared groups' scores and (b) will lead to evaluation results whose cause cannot be determined. If any of the elements are compromised and the effects of compromising are not clear from our discussions, we recommend consulting statisticians to explore alternatives and their effects on the prospective validity of inferences from results. As an alternative to consulting statisticians, we invite readers to call us.

III

Ratings

All US military services use ratings for evaluation. Ratings provide data for Officer Efficiency Reports (OER) and NCOER, for combat-attrition and other modeling, for scoring tactics and maneuver at the Army's Combat Training Centers, and for projecting readiness. Ratings also are used in training effectiveness evaluations, such as effectiveness evaluations of combat and maneuver training with new Army training devices. Despite the widespread use of ratings and the importance of decisions made on their outcomes, the Army and civilian personnel responsible for activities such as those mentioned above do not routinely report psychometric properties, such as reliability, of their ratings. As noted in Elementary Rule 10, the importance of estimating the reliability of scores lies, not only in implications for statistical power (Elementary Rule 3), but also in implications of reliability for estimating validity (Elementary Rule 11). These facts are routinely ignored in military evaluations of the kinds mentioned above. Any result of such ratings-based evaluations is therefore moot.

Scope

Two kinds of ratings are addressed here: analytic ratings and performance-appraisal ratings.[1] Analytic ratings are the kinds used by Burnside (1990), by Drucker and Campshure (1990), by Sherikon (1995), and in various user tests of new training. These kinds of ratings typically use multi-point rating scales on which SMEs estimate the extent to which

[1]These two kinds of ratings are subsumed by the methods Simpson (2000) called analysis and judgment.

proposed new training devices permit practicing or otherwise promoting learning of military tasks. Sherikon's (1995) rating scales, for example, used a value of zero to indicate tasks "not at all supported" by practice with CCTT, to a value of four to indicate tasks "fully supported" by practice with CCTT.[2]

Performance-appraisal scales are similar to training-capabilities scales or to device-capabilities scales but are applied to soldiers' collective or individual performance rather than to training or device capabilities. Observer-controllers at the Combat Training Centers, for example, rate units' collective performance on tactical tasks as T, P, or U, which stand for trained, partially trained, or untrained. Similar scales are used for individual performance-appraisal purposes, including the OER and NCOER mentioned earlier.

Misconceptions About Ratings

Field-trial results are more likely than ratings to be accepted as bases for approving new-training acquisition in the US Army. The reasons for this seem grounded in conventional wisdom that holds results from field trials to be inherently more valid than results from ratings. Judgmental overtones in our language support the conventional wisdom: Results of field trials are said to be objective and based on facts; ratings are said to be subjective and based on opinion. Evidence does not support the conventional wisdom: Many field-trial scores are based on opinions of judges serving in the capacity of observer-controllers, and ratings may, because of their reliability, yield inferences more valid than inferences from field trials.[3] Essential characteristics

[2]Appendix C is a summary of Sherikon's evaluation in which we re-analyzed results to estimate the reliability of Sherikon's ratings and the validity of inferences from those ratings.

[3]Compare, for example, the reliability of Sherikon's ratings in Appendix C to the reliability of live-fire tank-gunnery scores:

of data quality, that is, reliability and validity, seem not to be a part of the conventional wisdom.[4] Consider also:

(a) The reliability and therefore the maximum possible validity of SME ratings or of field-trial results do not inhere in whether field trials or ratings are used, but instead result from such factors as characteristics of data-collection instruments, control of data-collection procedures, numbers and kinds of observations, and the rigor and appropriateness of analytic methods.

(b) Any one-shot training evaluation, including an Army field trial such as IOT&E, is a point estimate. Because costs preclude replication, we know nothing about the generality of the evaluation result. Any result, including a null result, is likely to be no better than guessing.

(c) The farther we move along the continuum from tightly controlled, standardized classroom tests to multi-echelon, combined-arms field trials, the less will be the possibility for standardized training and test administration, and the less therefore will be the likely reliability of scores, their statistical validity, and the validity of inferences from those scores.

(d) Given equal numbers of field trials and ratings, reliability and therefore validity are likely to be greater for ratings than for field trials; this is so because administration is likely to be more standardized and error variance therefore less with ratings than with field trials.

Powers et al. (1975), on analyzing their live-fire tank-gunnery scores, concluded the reliability of those scores was no better than "random guessing" (p. 26). Inferences from those live-fire scores therefore could not be valid.

[4]To the best of our knowledge, there have been no Army training evaluations that permitted direct comparisons of the reliability and validity of SME ratings to the reliability and validity of field-trial results.

(e) Reliability and therefore maximum possible validity increase with increased numbers of observations,[5] and the costs of ratings usually are less than the costs of field trials. The bang for buck for any potential validity increment will therefore be greater for ratings than for field trials.

(f) Meeting the statistical-power requirements and neutralizing all the threats to valid inference that inhere in two-group, combined-arms, multi-echelon field trials are impossible. These threats bias such trials toward finding no statistically significant, and therefore no practically useful, differences between compared units' proficiency.

Advantages of Ratings[6,7]

As noted earlier, ratings often are used analytically, to assess training program or device capabilities. The obvious advantage of using ratings this way is that they lead directly to recommendations for increasing training capabilities. Additional advantages of ratings used analytically include:

(a) The results of analytic evaluations can be used to help justify budgets for training upgrades, by identifying tasks and mission segments that are unlikely to be trainable with the existing training program or device.

(b) Analytic evaluations are essential in designing multi-media training strategies, because analytic evaluations identify what cannot be taught with an existing program or device.

---

[5]Within broad limits.
[6]Material in this and the next section derives from a report by Boldovici and Bessemer (1994).
[7]See Grotte, Anderson, and Robinson (1990) for advantages and disadvantages of judgmental methods in defense analyses.

(c) Analytic evaluations can be performed with specifications and mock-ups, before programs are cast in concrete or metal is bent.

(d) The price of analytic evaluations is small compared to the price of field trials.

(e) Analytic evaluations are useful in forming hypotheses and questions for empirical investigation: What effect, for example, will CCTT's inability to support various lower-echelon tasks and mission segments have on the transfer scores of fully qualified crews, platoons, and company-teams?

Disadvantages of Ratings

In contrast to the advantages noted above, ratings used analytically have three disadvantages.

(a) The information from analytic evaluations yields weaker inferences about the effects of training than do the inferences that ensue from tightly controlled empirical evaluations: The results of analytic evaluations applied to date have been unsuccessful in estimating transfer.[8,9]

(b) The persons who perform analytic evaluations must be familiar enough with parent weapons systems, field operations, and mission S.O.P. to be able to identify subtle differences between training capabilities and field practice. The persons performing analytic evaluations also must have enough human-learning expertise to be able to make tenable inferences about the transfer effects of similarities and differences between training devices

[8]See, for example, Morrison and Hoffman's (1992) results of applying Pfeiffer and Horey's (1988) ratings-based transfer-estimation scheme.
[9]The same is, and will continue to be, true for two-group, large-scale, multi-echelon, combined arms field trials of Army training.

and parent weapons systems. The number of persons who combine these capabilities is small.

(c) The results of analytic evaluations derive from analysts' expertise rather than from so-called hard data. As already noted, conventional wisdom erroneously holds field-trial results inherently more valid than ratings. The conventional wisdom requires ignoring the reliability and validity considerations discussed earlier.

Essential Properties of Ratings

Essential properties of ratings are the same as essential properties of other scores, including field-trial scores. The essential properties are reliability, validity, and generality,[10] which provide the basis for our first two Rating Rules.

1. *Estimate and report inter-rater reliability and its implications for validity.*

Estimating the reliability of ratings is done by computing agreement between or among raters. As is the case for all other scores, the reliability of ratings is important for estimating consistency and increasing statistical power. Recall that reliability places statistical limits on the validity of all scores, including ratings, and by extension on the validity of inferences one may legitimately make from results based on any scores. High reliability never guarantees high validity. Rater agreement might, for example, be 100% wrong. Low reliability, however, does guarantee low validity. Correlation and other methods appropriate for estimating inter-rater reliability are in Appendix C. Appendix C also includes scratch-pad methods for estimating inter-rater reliability and methods for using the inter-rater reliability estimates to estimate maximum validity.

[10]See Elementary Rules 10 and 16.

2. Estimate and report generality.

The need for estimating the generality of scores based on ratings is, as is the need for estimating the generality of any scores, self-evident: We should like to know the extent to which our one-shot, point estimate – whether from ratings, field trials, or other evaluations – is representative of results we might expect from large numbers of replications of our one-shot point estimate. We know of no simple, scratch-pad methods for estimating generality. Statisticians, special software,[11] or both are required.[12]

Designing Ratings for Reliability

Designing ratings for reliability reduces to an exercise in unambiguous communication. Reliability will increase with the extent raters agree on the meaning of our instructions and scales. Lopez (1998) said it nicely: "A rating scale is an aid to disciplined dialogue."

The rating process may be conceived as having three phases: (1) Rater Preparation, (2) Observation, and (3) Recording.

Phase I: Rater Preparation.

The reliability of ratings will increase with the uniformity of understanding among raters about the rules of observing, rating, and recording. Raters should be standardized, and we should take measures to assess the extent to which our attempts to standardize raters have succeeded. Rating Rules 3 through 6 apply:

---

[11]See, for example, *Resampling Stats*, stats@resample.com and www.resample.com.

[12]Additional discussion of generality is in Elementary Rule 16.

3. *Be specific in instructions to raters.*

Reliability will be greater when the instructions to raters are specific than when instructions are general or loosely stated.

4. *Provide instructions early enough to allow practice, feedback, and learning.*

Do not give instructions so far in advance of ratings as to permit forgetting or so late as to preclude practice, feedback, and learning.

5. *Provide practice in observing and rating.*

Provide practice and feedback for raters, and elicit their comments on reasons for disagreements. Attend particularly to "I-thought-you-meants." Revise and re-try instructions and scales accordingly.

6. *Test raters.*

Use tests, that is, practice ratings, to make sure raters are capable of making the observations and distinctions you want. Estimate the inter-rater reliability of test ratings. Improve inter-rater reliability as suggested in Rating Rule 5 and in Elementary Rule 10.

Phase II: Observation.

Even with careful rater preparation and totally standardized raters, reliability will be affected by variables at work during observation of the characteristics or events to be rated. Rating Rules 7 through 16 apply:

7. Deconstruct multi-dimensional criteria.

Unidimensional criteria are more likely to yield reliable ratings than are multidimensional criteria. An example of bad practice here is an item that judges are asked to rate acceptable or unacceptable on a research-proposal evaluation form: "The proposed effort does not duplicate existing science and technology efforts elsewhere, and offers a unique or complementary approach to fulfilling an Army need" (Anonymous, undated). The item contains the multidimensional criterion of nonduplication, and uniqueness or complementary, and the accompanying instructions say nothing about how we should adjust our rating if, for example, the proposed "effort" duplicates only one other effort, rather than two or more "efforts" as given in the criterion. The criteria in the sample item need to be deconstructed, with separate scores applied to each.

8. Deconstruct multi-dimensional events.

Avoid time-sharing in rated events. Reliability is more likely to result from making single or small numbers of observations rather than large numbers of simultaneous judgments. Asking judges to provide a summary task performance rating based on observing component subtasks, for example, is not a good idea. More reliable ratings will accrue when component subtasks are rated separately.

9. Make transient events stable.

The reason for using still photography, stop-action video photography, and other instrumentation for scoring live-fire tank gunnery is to impart stability to events that are naturally transient – for example, a 120 mm. target-practice round passing through a plywood target at 1200 m. from the observer. Scoring transient events made stable by

photography and other instrumentation will yield more reliable results than will scoring the same events as they are happening.

10. Avoid noise in rated events.

Noise is not necessarily auditory. In ratings of soldiers' or units' performance in field trials, noise is more likely to be visual. An example of unwanted noise effects is in Kuma and McConville's (1982) report of the UCOFT Operational Test: "Dust completely obscured the target after the first tank fired. This condition became increasingly worse throughout the post-training BFD [battlefield diagnostic] test" (p. C-10). It is hard to imagine how observer-controllers scored hits and misses for "completely obscured" targets, or for that matter, exactly how the condition could become "increasingly worse" from a scorer's (i.e., rater's) standpoint.

11. Strive for observability in rated events.

An implication for training-capabilities ratings is that reliability will increase with the extent to which raters may access the rated training device to provide reality checks on ruminating about what the device does or does not teach. Training-capabilities ratings are more likely to be reliable and valid when the training and attendant simulations are accessible to raters than when they are not.

12. Require comparative rather than absolute judgments.

Paired-comparison and partial paired-comparison techniques require judgments of more or less and produce high inter-rater reliability. Their use is tedious but worth the effort. Adaptations of McCormick and Bachus's (1952) and of Rambo's (1959) design guidelines are in Boldovici, Harris,

Osborn, and Heinecke's (1977) evaluations of tank-target threats.

13. Alert raters to likely errors.

Analyze the results of rater practice (Rating Rule 5) and rater testing (Rating Rule 6) to identify common errors. If common errors cannot be eliminated by revising instructions and scales, then alert raters to conditions under which the errors are likely to occur.

14. Allow raters to observe and rate more than once.

Here again, our concern is with reliability and its implications for validity, statistical power, and generality. Multiple ratings of the same event will, within broad limits, yield more reliable scores than will single ratings.

15. Provide scoring aids or templates.

The benefit of templates for reliable scoring can be seen by imagining the task of assessing an umpire-trainee's strike-calling proficiency. Three templates are involved here, as given by (1) the rule-book definition of a strike zone, (2) the umpire-trainee's image of a strike zone, and (3) the image of a strike zone held by the person scoring the umpire's performance. To the extent the rule-book template can be replicated (as with photo-electric beams, for example),[13] the measurement error introduced by the scorer decreases, with salutary effects on measurement reliability and on the validity of inferences from the scores.

[13]The beam-and-buzzer system for signaling out of bounds in tennis is, of course, a similar example. Less so is a football coach's prerogative to request referees' reviews of called infractions.

Fruitful Army applications of such systems might include, for example, scoring the accuracy of tank gunners' sight pictures and, thanks to the computing power of modern AAR-support systems, the adequacy of units' distribution of fire in field exercises. Application of the thinking here also can be seen for so-called soft skills, as in the test items produced by Project Alpha to rate junior leaders' management skills: Checklists rather than pictorial templates are used in this case as aids to rating, for example, the adequacy of an NCO's probes or follow-up questions in response to a subordinate's request for time off to visit his pregnant girlfriend.[14]

16. Do not require raters to process results.

Asking raters to provide summary scores, percentage of target hits for example, from individual scores invites errors and therefore unreliability in ratings. Generating the summary scores from raters' component scores is better left to the test organization.

Phase III: Recording.

Even with adequate observer preparation and careful control of the rating process, reliability will be affected by variables operating during the recording of scores. Two such variables – timing and simplicity – give us Rules 17 and 18:

17. Keep the time short between observing and recording.

Reliability will increase with decreased time between observation of the event or characteristic of interest on the one hand and recording the rating on the other.

---

[14]We thank Bill Osborn and Larry Meliza for suggesting the examples used here.

18. Keep the rating forms simple.

Reliability-prone forms minimize the amount of judgment and decision-making required for their use. Simplicity in rating forms reduces data-recording time and should allow more time for observers to observe.

Additional Sources

The Questionnaire Construction Manual (Babbitt & Nystrom, 1989) contains a wealth of information for writing instructions and designing forms for reliability.

The American Institutes for Research in Georgetown has a Document Design Center staffed by persons experienced in subtleties of English use for clarity of instructions.

Grotte, Anderson, and Robinson (1990) discuss a variety of judgmental methods and the conditions attending their utility.

IV

Suggestions

1. We probably should try something different.

Our attempts to demonstrate transfer of higher-echelon, device-based training to field settings have, as noted in earlier chapters, failed. This fact suggests we should explore alternatives to the thinking and methods that attended our failures. We therefore present some alternative thinking and methods here – by no means the last or the best word on training evaluation – but in hope of sowing a seed or two from which new evaluation cultures may grow. Several essential orientations seem desirable; they provide bases for our Suggestions 2 through 6.

2. Meeting the training-evaluation challenge is in some ways analogous to successfully conducting a hasty attack.

The operative word here is hasty, a modifier that does not characterize training-evaluation practice in the US Army.[1] Chief points of the analogy are: (a) knowing what a target of opportunity looks like, (b) having G2 capable of informing us what the targets of opportunity will be doing before the attack, and (c) arming shooters with an arsenal capable of coping with most likely targets of opportunity.[2] The need for commanders experienced in distinguishing good

[1]Plans for the training-effectiveness portion of the Independent Operational Test and Evaluation for the Close Combat Tactical Trainer, for example, were over 6 years in the making.

[2]Concomitant abilities include being able to recognize when the window of opportunity has closed for a target, to correctly assign priorities to targets, and to recognize when conditions contraindicate attacking.

plans of attack from those likely to fail goes without saying.

At the risk of belaboring the analogy, the implications of our points for improved training evaluations are: (a) knowing how and where units train and how they are likely to use new training, (b) total familiarity with production and delivery schedules for new training and with extant data streams, and (c) equipping personnel with tool kits full of evaluation methods, including methods for tapping extant data streams, to be used singly or in combination for achieving the Army's training-evaluation objectives. The need for commanders capable of distinguishing good evaluation plans from likely failures goes without saying.

Evaluation success requires the three elements just presented and the command structure to be in place.[3] To the extent any of the elements or pieces of the elements are missing, as they were in earlier forays, we preempt our chances of meeting the training-evaluation challenge.

3. *The complexity of higher-echelon, device-based training guarantees that any single index of effectiveness will be meaningless.*[4]

Summative evaluations, which by definition have little diagnostic value, continue to drain resources better used for diagnostic evaluations. Evaluating the effectiveness of complex, higher-echelon training systems poses a challenge that exceeds the challenges in previous evaluations. We have entered

[3]A training-evaluation command structure comprising the elements and capabilities outlined here does not exist. An attempt to remedy that situation is in the consortium of organizations – TRADOC, ARI, OEC, TRAC, PM CATT and others – whose representatives are exploring ways to institutionalize long-term evaluations of CCTT-based training.

[4]Variations on this theme are in Appendix C.

a training era dominated by Distributed Interactive Simulation (DIS) worlds and evolving into interoperable federations of multiple worlds bound together in a High-Level Architecture (HLA). Both DIS and HLA simulations must by design serve needs of many training customers simultaneously. With such great numbers and kinds of components, of virtual environments, and of customers, any thinking that DIS and HLA systems can succeed entirely or fail entirely begs for dismissal as nonsense: Some parts and functions will work well and will satisfy customers; other parts and functions will work less well and will not satisfy customers. Because higher-echelon, device-based training systems are multi-component and interactive, it makes no sense to try to measure either the effectiveness of the total system or some composite of the effectiveness of separate system components.[5] System-level total measures do not highlight needed improvements in parts and functions. Our current ignorance of the interaction effects among components of complex training systems at the same time precludes attempts to combine measures in meaningful composites.[6] Furthermore, total system-level or composite measures of effectiveness require validation by more than opinion. Development and validation of such measures is a legitimate endeavor, but improving military training cannot wait for the results.

---

[5]Software-engineering orientations, comprising continual cycles of tryouts and revisions, would better serve our ends than do existing summative, one-shot, go/no-go (e.g., IOT&E-like), training-evaluation orientations.

[6]The analytic dilemma here is perhaps nowhere closer to home than by analogy with medical science's ignorance of drug interactions: Take a few daily prescriptions for high blood pressure, an aspirin a day to "thin" your blood, calcium supplements to delay osteoporosis, garlic tablets to counter cholesterol, four or five vitamins whose benefits are touted in the popular press. The numbers of interactions quickly get into millions – numbers which preclude knowing what the interactions are.

4. *Higher-echelon, device-based training must be evaluated in systems terms.*

As implied in Suggestion 3, the value of modern, device-based Army training 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.[7] Consider for example the recently fielded CCTT, which provides the core medium for simulated maneuver training in the Army's Combined Arms Training Strategy (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 is on the CCTT's contribution to the mix: What part of the total CATS burden should we be asking the CCTT to bear? Evaluations should address how the CCTT and forthcoming devices (members of the "CATT family" for example) complement or supplement existing training alternatives to support and implement CATS while remaining within current and future budgetary limits. The evaluation scope must consider all consequences and side effects on the total Army training system. Analytic studies without experimentation can derive estimates of CCTT impacts. Evaluations must also allow the conclusion that CATS existing at any given time needs to be revised.

5. *Approach evaluation of modern, device-based, higher-echelon military training as part of a larger evaluation program applied to the total Army training system and directed toward continual training improvements.*

---

[7]A reviewer commented that our focus in an earlier report (Boldovici & Bessemer, 1994) on device-based training's relation to the total Army was too narrow: "[Consider] evaluation also as applied to joint and combined operations . . . get other Services and some potential Allies involved" (Frederic J. Brown, personal communication, 17 January 1994).

This is not a new idea. In his summary of a program-evaluation symposium (American Institutes for Research, 1970), Baxter wrote,

> Too often evaluation has been a tag end in the later phases of a project. The desired alternative is to make it a systematic ongoing process beginning with the planning phase (setting goals) through the design of the program activity [new Army training in our case], to the collection and interpretation of outcome data. Evaluation should be a continual process because its findings can serve to modify goals and help to redesign certain aspects of the program. It sets up a repeating cycle for improvement (p. 159).

Later elaborations of Baxter's theme were called Total Quality Management (TQM).

A reasonable assumption is that at the outset little will be perfect about initial versions of new device-based, higher-echelon training and their use. Accepting the premise of imperfection, we advise reforming acquisition and implementation processes to encourage frequently repeated upgrading and improvement of training. One of the advantages of DIS and HLA modular architecture is that they facilitate systematic, step-by-step, module-by-module improvement. Continual feedback from analytic evaluation results is the essential ingredient in defining requirements for modifications and for assessing the success of each modification.

6. *Evaluation policies and processes for higher-echelon, device-based Army training programs must be planned as continual, institutionalized*

parts of a superordinate, total-Army, TQM-like system such as mentioned earlier.[8]

This suggestion was implied in the fifth suggestion: Evaluations of new training must be planned as continual, institutionalized parts of an over-arching, TQM, total-system process. Although an Army-wide TQM program for training does not exist, some believe there will have to be one[9] or something like it if the Army is to be capable of maintaining readiness. Collection and analysis of training input, process, and outcome data must become integrated with all Army training to enable continual training quality improvement and a modicum of quality assurance.

The organization of the Army Center for Lessons Learned (CALL) and its support of the Combat Training Centers provide a model for institutionalizing a TQM-like system to support new Army training.[10] Practices of continual evaluation have become institutionalized at the CTCs; building in capabilities for evaluation to all new Army training will help move other parts of the Army training system in the direction of continual, institutionalized evaluation.

The Arsenal

To continue our analogy with shooters armed with weapons appropriate for ambushing targets of opportunity,[11] descriptions of several such weapons

---

[8]See Bessemer and Myers (1998) for specific suggestions related to the tactical and maneuver training.

[9]See, for example, Booher and Fender (1990).

[10]Diana O. Tierney (personal communication, December, 2000) noted that for CALL, "to be truly systematic . . . there needs to be consolidated feedback to a higher order organization, office, or agency responsible for fixing the entire 'training factory' (e.g., the agency that writes collective training policy and doctrine) and to the agencies responsible for writing the training support packages sent to the units and for training the individual soldiers sent to the units."

[11]So as not to be misunderstood – we know the shooters and other persons who meet the requirements for all the essential ingredients

follow.[12] The methods we suggest are not exclusive alternatives, but rather are means to be used at one time or another to address evaluation questions for which they are suited. No single method serves all ends.[13] Use the methods in combination to provide converging evidence compensating for the weaknesses of each. The methods are:

1. In-device learning experiments.
2. Quasi-transfer experiments.
3. Correlational research with archived data.
4. Efficient experimental designs, for example, randomized block, Latin square, and analysis of covariance (ANCOVA).
5. Quasi-experimental designs.
6. Analytic evaluations.
7. Improved methods for documenting training.

1. *Conduct in-device learning experiments to examine the effects of altering training conditions.*

Many hypotheses about what to train, how to train, and how much to train can be tested without incurring the expense of field exercises. In-device learning experiments are usable with groups of similar units that obtain device-based training using

---

outlined earlier. Those persons are few, and they do not all go to work under the same roof. The roof – that is, an effective organizational structure – has not yet been built. We hope the suggestions in this chapter will help Army leaders build the roof and seek out the shooters.

[12]Our examples deal mainly with simulator- or device-based training because that is where the majority of our recent experience lies. Most of the information in the examples is, however, applicable to training other than device-based.

[13]Invalid inferences about training effects are just as likely from these evaluation methods, perhaps more so because of their complexity, as are invalid inferences from the methods we have used to date. Identifying experts, doing evaluations in accordance with their advice, and attending to their views on valid inferences from evaluation results, are imperative. Many such experts are bio-medical statisticians.

the same set of exercises or very similar exercises. Many National Guard tank platoons, for example, perform the same exercises in their initial Virtual Training Program weekend in SIMNET at Fort Knox. These units become candidates for an evaluation when they train at different times in an interval when the device has no material changes. With units randomly assigned to treatments, the evaluation continues over time until each treatment group has enough units to obtain the needed statistical power to enable valid inferences about treatment differences. Opportunities for comparable in-device learning experiments may soon emerge at CCTT training sites as units routinely select similar exercises from a standardized library.

The measures used as dependent variables for in-device experiments may derive from performance of tasks repeated in successive exercises (e.g., calls for indirect fire and adjustment of indirect fire), or from general outcomes that apply to all exercises (e.g., unit losses). Such measures directly quantify learning as improvement across exercises and enable statistical comparison of amounts of learning between treatments. Accuracy of calls for fire, for example, or survivability of unit vehicles might improve at different rates with different treatments. These kinds of measures enable repeated-measure experimental designs that provide statistically powerful comparisons with small samples, because the pre-existing differences among units do not affect differences in rates of change between or among treatments.

The in-device learning experiment is a good way to evaluate device improvements, such as a multimedia after action review (AAR) support system expected to improve learning in the AAR. During the in-device experiment, units assigned to an experimental group use the AAR system, while units in a control group do not. The experimental hypothesis might be that

tasks reviewed with support by the AAR system will show greater performance gains in subsequent exercises than the same tasks reviewed without support.

When performance cannot be consistently measured in each successive exercise, in-device learning may be assessed using posttests, or pretest-posttest changes, derived from performance of tasks, subtasks, or other elements that were included in the training exercises: A set of eight training exercises might, for example, include "support-by fire" and "conduct an assault" tasks only twice each. Evaluators can examine differences in learning these tasks between groups of units training with different treatment conditions (e.g., AAR systems, using measures obtained in a posttest exercise that includes both tasks). This approach also is useful to address questions about how much training needs to be given to attain specific levels of performance for particular tasks or groups of tasks. In this case, all units can train with the same sequence of exercises. Subgroups of units perform a test exercise inserted at different points in the sequence. Our analysis then examines the relation between performance measures and the number of task repetitions completed before the test.

A chief advantage of in-device learning evaluations is that their results are more likely than field trials to yield useful diagnostic information. This is so because of opportunities for tighter experimental control and for increased statistical power resulting from greater numbers of observations. Once we identify efficient ways of training via in-device learning experiments, we can investigate transfer of training in field trials with reduced risk of wasting money.

2. *Conduct quasi-transfer experiments to supplement, replace, or simulate field transfer experiments.*

Quasi-transfer experiments provide training with a simulator or other training device and assess transfer using a different device, new conditions in the same device, or new conditions in a reconfigured version of the same device. Opportunities for quasi-transfer experiments arise whenever the environmental and operational conditions of a simulated transfer test approximate a battle situation to which we want transfer of training to occur. Transfer performance measured in a quasi-transfer experiment yields an estimate predicting performance in the battle situation. In principle, this reasoning is no different than using performance in a field exercise that simulates a battle situation to estimate performance in the battle situation. In both cases, professional military judgment must assess the accuracy and validity of the estimates by considering how well the device or field simulation presents the conditions that influence performance in battle.

Quasi-transfer experiments provide low-cost evaluations that can yield valid empirical and analytic results. Examples are experiments by Witmer (1988) and by Turnage and Bliss (1990), who used quasi-transfer experiments in research with tank-gunnery training devices, and of Lintern (1987), who used quasi-transfer experiments in research with fixed-wing aircraft training devices. Experiments of this kind can replace or supplement field transfer experiments when a field exercise cannot simulate, or poorly simulates, our battle conditions of interest. For one example, simulation devices represent better than field exercises the effects of indirect fire. For some tasks, such as "React to Indirect Fire," device-based test exercises may be the only or the best choice for estimating

transfer. In other cases, a simulated pilot-test of a field-transfer experiment can help refine our experimental design. Trying out experimental procedures with simulated conditions and device-based performance measures can help us avoid mistakes or omissions that will invalidate the results of expensive field trials.

As an example of the utility of quasi-transfer experiments in Army training, suppose a US Army, Europe (USAREUR) commander questions what kind of tank-appended weapon simulator units should use for engagement simulation in field exercises. His concern is that the current laser device does not employ precision gunnery techniques and may interfere with the retention and sustainment of gunnery skills. He wants to know if USAREUR should acquire a more costly device tied into the tank fire control system enabling crewmen to perform all standard gunnery procedures, including precision gunnery techniques. A quasi-experiment is the obvious choice to compare the effects of the two devices on gunnery skills. After each unit in one of two device-groups completes field exercises with its assigned device, the commander tests the performance of gunners and tank commanders in a gunnery simulation device such as the Unit Conduct-of-Fire Trainer. Because of ammunition costs, a live-fire field test could assess only performance with an extremely limited number of rounds and with only inadequate numbers of crewmen. The quasi-experiment, on the other hand, allows for repeated commander and gunner performance measures, thereby providing the reliable performance estimates our commander needs to guard against making invalid inferences from his test results.

The utility of quasi-transfer experiments for new CCTT-based training also is apparent: Quasi-transfer experiments for CCTT might include training given

tasks and missions on CCTT and testing for transfer to new tasks and missions on CCTT. After training with certain conditions, such evaluations might assess transfer under different conditions. Commanders at posts with CCTT sites, such as Fort Hood and Fort Stewart, undoubtedly are interested in what policies they should establish for CCTT use: How, for example, should I allocate CCTT time between platoon- and company-level training? Quasi-experiments are obvious candidates for examining the amount and kind of CCTT platoon training needed to make company-level training sufficiently effective or more so. After various platoon training trials in CCTT, the degree of improvement of learning in company-level CCTT training exercises would measure the transfer effects produced by the various numbers of trials.

Additional examples of using quasi-transfer experiments to assess the utility of CCTT are in Appendix D.

The chief advantage of quasi-transfer experiments is in using them for estimating 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 inferences (see Introduction and Elementary Rules 10 and 11), 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 inference than are the smaller numbers of scores

typically available from field trials.[14] 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 CCTT only partially supports?

3. *Conduct correlational research with archived data.*

Collecting and storing training data and related information as part of a TQM system at simulation-based training facilities and field training sites will, we hope, eventually become a priority for training developers and evaluators. The CCTT facility at Fort Hood, for example, could obtain and accumulate data of several kinds, including: (a) how trainers and units prepare for CCTT training, (b) what CCTT training was done and how it was done, (c) selected training performance data, and (d) surveys of trainers and unit personnel addressing customer satisfaction. Bessemer and Myers (1998) identified a number of indicators in each of these categories that should prove useful for continuing and accelerating the effectiveness of new Army training. The immediate payoff is an ability to monitor training quality and results over time to detect positive or negative trends and to detect warnings of potential problems.

The longer-term, and in our view, greater payoff comes after a considerable body of data accumulates over time – six months to a year depending on the frequency and intensity of utilization at the facility or site. Analysts then can apply various inferential methods to the amassed data. Unit training outcomes, for example, might be

[14]For a discussion of reliability and validity, see Appendix C.

categorized with a 3-5 point ordinal scale from good to moderate to poor. Analyses can identify variables that predict unit membership in the outcome categories. Multiple regression methods can identify predictors of quantitative variables, such as customer satisfaction ratings. Such findings contribute to building up a picture of training best practices for new training such as CCTT. As is the case for all correlational research, the results do not prove that certain variables cause the values of the predicted variable. The findings do, however, suggest causal hypotheses testable by additional evaluations.

As sample sizes increase and more data accumulate over time, our ability to discover complex and subtle effects of various training variables increases, as long as the training management and practices remain stable. And if major changes are made, the pattern of relationships existing before and after the changes can be compared. Differences associated with the pattern of these relations will infuse objectivity into assessing the effects of changing training management and practices.

4. *Use efficient experimental designs to control sources of variation and thereby increase statistical power.*[15]

Recall that a common reason, perhaps the most common reason, for null results and invalid inferences about equal effectiveness of the kinds of training we wish to compare is inadequate statistical power: Evaluations with few numbers of observations, or with unreliable scores, or with various other flaws preclude finding statistically significant differences between our compared

[15]The efficient experimental designs outlined here require some statistical expertise in planning their use and in interpreting their results. We recommend consultation with experts before deciding to use these designs.

groups' scores. To the extent we control sources of variation that affect evaluation outcomes, the statistical power of our evaluations, the precision of our estimates, the validity of scores, and the validity of our inferences from evaluation scores increase.

Unlike commonly used experimental designs such as *t*-tests and analyses of variance (ANOVA), the efficient designs discussed here control unwanted sources of variance and have salutary effects on power, precision, and validity without increasing sample sizes.

We discuss the utility of two kinds of efficient designs in Sections 4.1 and 4.2: Randomized block designs and Latin square designs.

In Section 4.3, we discuss a third method of increasing the efficiency of experimental designs: the analysis of covariance. ANCOVA is a correlational method that may be used in conjunction with, or as an alternative to, other evaluation designs.

In addition to these basic designs, there are designs that are more complex, built up from variations on our basic designs but using similar principles.[16]

4.1 *Use randomized block designs when your prospective sampling elements, such as battalions, companies, platoons, crews, squads, or individuals, form natural groups.*

[16]Keppel (1991), Kirk (1995), Myers (1979), and Winer, Michels, and Brown (1991) described complex designs in texts written for behavioral research practitioners. Other evaluation-design texts for applied industrial, agricultural, or biomedical research may provide useful insights for military-training evaluations but demand caution in their use, because they may omit considerations that uniquely attend human-performance evaluations.

In randomized block designs, we assign "natural groups" randomly and assign natural sub-groups randomly to treatments, for example, new training and old training. The groups are the "blocks." This arrangement removes differences among groups from the comparison, while retaining sub-group variation. Thus may we increase statistical power, as compared to the power of the fully randomized *t*-test or the ANOVA.

Natural groups amenable to blocking in Army training evaluations are obvious: Battalions comprise companies, companies comprise platoons, platoons comprise crews, crews comprise squads, and squads comprise soldiers. Why risk confounding battalion-, company-, platoon-, crew-, or any other echelon effect with treatment (kinds of training in our case) effects when we may avoid such confounding by distributing echelon effects across treatment effects – especially when an evaluation design, namely randomized blocks, removes differences between groups in echelons and thereby yields the concomitant benefits of (1) increased statistical power, and (2) increased potential for valid inferences from evaluation results? (Answers along the lines of administrative convenience and adherence to unexamined tradition come to mind.)

Table IV-1 is an example of a randomized-block design. Characteristics of this design are (1) it comprises two treatments (kinds of training) and a no-training control group, and (2) the treatment effects resist confounding by echelon effects, because each echelon gets all treatments.

Table IV-1
Example of a Randomized Block Design.

	Table VII Practice Conditions		
Unit	Live-Fire	Laser	None
Bn 1, Co A	Plt 3	Plt 1	Plt 2
Bn 1, Co B	Plt 2	Plt 1	Plt 3
Bn 1, Co C	Plt 1	Plt 2	Plt 3
Bn 2, Co A	Plt 3	Plt 2	Plt 1
Bn 2, Co B	Plt 1	Plt 3	Plt 2
Bn 2, Co C	Plt 2	Plt 1	Plt 3
Bn 3, Co A	Plt 1	Plt 2	Plt 3
Bn 3, Co B	Plt 2	Plt 3	Plt 1
Bn 3, Co C	Plt 3	Plt 1	Plt 2

This randomized-block design well serves the commander who seeks objective answers to his questions about, for example, whether a vehicle-appended laser simulator device can substitute for part of the practice tank crews normally do with live, target-practice, ammunition. The commander views Table IV-1 and the inferences possible therefrom in the following light: I have nine companies in three battalions with three platoons in each battalion. Each of the three platoons is randomly assigned to one of two treatments (live-fire or laser) or a control group (none). The no-training control group provides a baseline that I shall use to gauge the benefit of live-fire and laser practice.[17]

[17]Our commander, anticipating reactions to his never-before-considered evaluation design, speculates about inevitable objections to denying practice to members of his no-training control groups. He concludes, "All such objections are gratuitous: I can allow platoons in my no-training control groups to practice the omitted Table VII and refire Table VIII for record after my evaluation is done."

Our commander now conducts the thought experiment to examine how our Table IV-1 design can answer his questions: All platoons fire tank-gunnery Table VIII as usual, with crew scores used as the measure of performance. With one platoon from each company assigned to each treatment, any difference in average performance between companies is unlikely to affect the difference due to treatments. Any differences between the post-training scores of his compared groups are therefore far less likely to be due to variations among platoons within companies or to crews within platoons than to the compared kinds of training.[18]

Additional examples of blocking variables that can be used to advantage in Army training evaluations are:

(1) Trainers, to control for variations in how trainers conduct exercises or after-action reviews.
(2) Time of day, to control for variance in performance due to sun angles and other determinants of visibility.
(3) Geography, to control for possible effects of terrain on compared groups' performance.
(4) Pretest scores, to match groups in terms of pre-training proficiency as estimated by the dependent variables.

[18]An esotericum: With our Table IV-1, randomized-block design, we have four crews per platoon (known but unshown) and therefore four measures of performance within each platoon that provide independent estimates of variance among crews. These estimates of crew variation can be combined with the estimate of platoon variations: The required sample size thus becomes the number of crews rather than the number of platoons, giving us an automatic increase in statistical power and therefore a greater likelihood of valid inferences from evaluation results. Such freebees abound, thanks to the existence of echelon hierarchies. Competent evaluation planners reveal themselves via knee-jerk reactions that parlay these freebees into no-cost increases in statistical power and attendant validity of inferences from evaluation results.

(5) Observers, to control for individual differences in the stringency with which they rate performance.
(6) Mental ability categories as given by the Armed Services Vocational Aptitude Battery's (ASVAB) classifications.[19]

4.2. *Use Latin square designs to control for the effects of two blocking variables.*

Latin squares are similar to randomized block designs, because a blocking variable is used to group sampling elements. Latin squares, however, impose a second constraint on the arrangement of sampling units. This constraint comes in either of two forms:

(1) The Latin square cross-classifies the blocking variable by a second natural grouping or measured blocking variable.

(2) The second ("cross-classifying") blocking element may represent, not a second natural grouping, but some manipulated independent variable that naturally attends the sampling units.

In this second case, the evaluator usually has little or no interest in the effect of the second variable; he sees it instead as a nuisance variable. But it is a nuisance that must be contended with: It cannot be held constant in the evaluation and, if not controlled, promises to confound the effects of compared kinds of training. An example is an evaluation done at

[19]Blocking on ASVAB-given mental-ability categories yields another freebee: In addition to increasing statistical power, blocking on ASVAB categories sets the stage for sage commanders' answering questions such as the following: Is the new training equally effective for all my squad leaders, regardless of their mental category? Or do the more effective training conditions vary between the higher and lower mental categories?

terrain-exercise areas on five different Army posts. The five posts form the first-level, natural-group, blocks for our evaluation. Assume now that the sage commander has reason to believe that variance among observer teams in their scoring of exercise performance will affect evaluation outcomes. He therefore chooses as his second, cross-classifying, variable, five three-man observer teams that rate performance of one tank company completing a standard exercise at each post. This cross-classified, Latin square, arrangement not only increases the statistical power of our training evaluation, but also controls for (1) the effects of terrain and (2) the effects of differences among observer teams in how they rate company performance.

The cross-classification sets up a square matrix of cells with rows defined by levels of one blocking variable and columns defined by levels of a second cross-classifying variable as summarized in Table IV-2.[20] The sampling elements in each cell combine one row-variable value with one column-variable value.[21]

---

[20]Ebony-tower residents, *n.b*: Our example, presented as hypothetical, is applicable to evaluating the Force XXI Battle Command Brigade and Below (FBCB2) system. Our example provides statistical-power advantages and control of nuisance variables that are impossible with customary, two-group, *t*-test evaluations. We welcome, as always, hearing views to the contrary.

[21]The counterbalanced arrangement of the Latin square, including our example in Table IV-2, combines each treatment condition once with each level of the row variable and once with each level of the column variable. The statistical analysis of this design estimates training effects from the data unaffected by either row (observer team) differences or column (location) differences. The analysis removes variation associated with both observer-team and location variables from the remaining random variation among treatments. And the smaller random difference among treatments increases the power of statistical tests relative to a corresponding randomized block design controlling only the column or row variable.

Table IV-2
Example of a Latin Square Design

	Observer Team				
Terrain	1	2	3	4	5
Location 1	A(2)	B(5)	C(4)	D(1)	E(3)
Location 2	B(1)	C(2)	A(5)	E(4)	D(3)
Location 3	C(2)	E(4)	D(5)	A(3)	B(1)
Location 4	D(5)	A(3)	E(4)	B(2)	C(1)
Location 5	E(4)	D(3)	B(1)	C(5)	A(2)

Note. Letters in each cell of the square denote the treatment conditions. The numbers in parentheses show one company at each location randomly assigned to each cell.

A special kind of Latin square is the repeated-measure Latin square, which takes its name from the repeated measurement of performance by each sampling element on successive occasions. Our discussion of appropriate conditions for use of repeated-measure Latin squares is in Appendix D, as are additional considerations affecting the use of Latin squares.

4.3. *Use analysis of covariance (ANCOVA) when measurable independent variables are present and are difficult or impossible to control or to vary.*

The measured independent variable in ANCOVA is a covariate, examples of which are pretest scores, platoon leaders' experiences, and target visibility. The ANCOVA becomes useful when the evaluator knows or suspects for good reason that one or more covariates have a strong effect on the dependent variable, for example, ratings of units' performance of a hasty attack, in the evaluation.

ANCOVA is an alternative to using the measured covariate to form matched groups in the randomized

block designs discussed earlier. The ANCOVA is more practical when the number of sampling units is too small to find enough units with equal covariate scores to form blocks. Another reason for using ANCOVA is that the covariate values may not be obtainable in advance, leaving no time to arrange units in blocks before assigning units to treatments.

The ANCOVA combines correlational methods with other kinds of experimental designs.[22] ANCOVA may involve one or several covariates; we discuss here only the case with a single covariate and single dependent variable.[23]

Two different sets of conditions motivate one's use of ANCOVA:

(1) The first set of conditions occurs when evaluators have little interest in the effect of a covariate but believe the covariate affects the dependent variable by increasing the amount of variation found among sampling units.

Example (1): No interest in covariate effects on the dependent variable(s) except that covariate affects dependent variable via variability among sampling units.

The increase in sampling variability produces an undesirable reduction in statistical power. In this instance, ANCOVA will remove a portion of the variability in the dependent measure that is correlated with variation in the covariate.[24] The

[22]The analyst must observe three cautions when using ANCOVA. These cautions are discussed in Appendix F.

[23]Repeated-measures designs such as the one described in the previous section also can employ ANCOVA methods. With repeated-measures designs, however, the covariate or covariates also must be measured repeatedly before each successive administration of the treatments.

[24]The variance removed is approximately proportional to r^2, so the dependent variance that remains is proportional to $1-r^2$.

reduced variation in the dependent measure then makes statistical estimates more precise and increases the power of statistical tests for differences between compared groups. The stronger the correlation (r) 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.

An example of this first kind of ANCOVA application is one with platoons randomly assigned to several simulator-based training conditions and then given a posttest exercise in the field. From past observations, we know that platoon leaders' experience levels influence the platoons' posttest scores. Our interest in the leaders' experience does not derive from curiosity about effects of this variable as an object of the evaluation, but from an expectation that it will increase the variation among the sampled platoons' posttest scores. The evaluators 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.

(2) The second set of conditions for using ANCOVA occurs when the effects of the covariate are a direct object of interest in our evaluation.

Example (2): We want to find out the effect(s) of our covariate on our dependent variables.

Randomized blocks with matching based on the covariate have the same effect.

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 a simulator pretest and field posttest scores, for example, in order to predict posttest scores. A linear prediction equation estimated from sample data could then 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 IV-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 for Group C shows that 1 day of training is insufficient regardless of the pretest score. The relation for 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. Statistical power accrues in this case because our estimates and inferences are based, not on the full variation of scores within groups, but on the much smaller variation among deviations of observed data points from the lines.[25]

[25]In contrast with in-device and quasi-transfer experiments, ANCOVA designs are more powerful, because in-device learning experiments and quasi-transfer experiments often use the same test to obtain pretest and posttest scores to measure change. The statistical analysis frequently uses gain scores, that is, difference scores obtained by subtracting the pretest from the posttest.

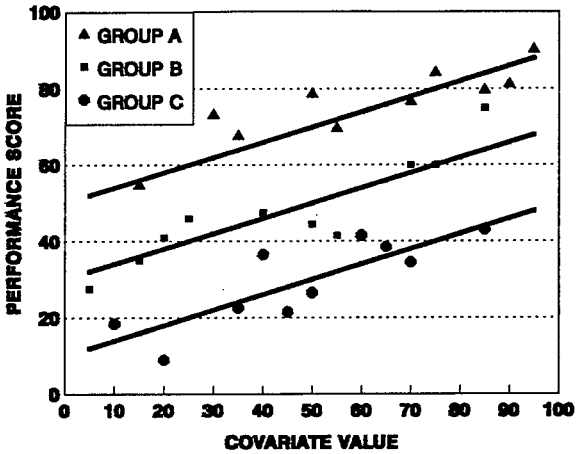


Figure IV-1. Example of parallel linear relationships within treatment groups in ANCOVA.

Appendix F contains additional considerations for using ANCOVA and an evaluation design for estimating lower-echelon effects on CCTT utility.

5. Use quasi-experimental design methods when controlled experiments with randomization are impractical or inefficient.

Quasi-experimental designs are peculiar arrangements of treatments, units, and occasions without complete randomization that nevertheless offer protection from particular threats to valid inference. When we can combine such designs with specific prior knowledge or supplementary data bearing on threats to valid inference, we may be able

---

Harris (1963), Reichardt (1979), and Venter and Maxwell (1999) showed that ANCOVA performed on the posttest scores with the pretest scores as covariate will be more powerful than the gain score analysis. The ANCOVA has the added advantage of revealing treatment-pretest interaction effects, that is, whether and how the treatment differences vary as a function of the pretest score.

to reach some valid inferences about training effects despite the absence of full experimental control.[26]

As an example of the utility of quasi-experiments for military training evaluations, Appendix G shows a quasi-experiment appropriate for evaluating transfer of new training to field performance. The opportunity to use this design arises whenever a major training system such as the CCTT is under development, and a schedule for fielding the system at various locations becomes available. Having units at locations served by newly fielded training that also are scheduled regularly for rotations at CTC sets the stage for evaluation with a quasi-experimental design, as shown in Appendix G.

6. *Evaluate training device capabilities analytically in all phases of the system life-cycle.*

Opportunities for analytic evaluations arise in all phases of device development: requirement definition, concept formulation, engineering development, system integration, and operational testing. After initial fielding and implementation, additional opportunities arise when product improvements or new subsystems are installed, or when operational doctrine, organizations, or task documentation affecting users is modified. An example is the addition of Future Battle Command Brigade and Below (FBCB2) capabilities to the CCTT at Fort Hood to support the 4th Infantry Division--the Army's first division conducting "digital" operations. Tasks performed by units operating with FBCB2 are substantially different from those performed by conventional units, extending CCTT training into a realm not covered in previous evaluations.

[26]Cook and Campbell (1979) discuss the kinds of variables evaluators must consider for several kinds of quasi-experimental designs.

Unlike empirical evaluations, analytic evaluations are based on experts' analyses and judgments 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 (1985) Device Effectiveness Forecasting Technique (DEFT), or Sticha, Singer, Blacksten, Morrison, and Cross's (1990) Optimization of Simulation Based Training (OSBATS). They also can use various methods reviewed by Knerr, Nadler, and Dowell (1984), including the paper-and-pencil checklist methods used by Burnside (1990) and by Drucker and Campshure (1990) to evaluate SIMNET's training capabilities. Other methods 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.

Summaries of Burnside's, Drucker and Camphsure's, and Sherikon's analytic methods, which led directly to tenable suggestions for improving new training, are in Appendix H.

Decision-making based on analytic evaluations becomes increasingly hazardous as time passes. Analytic results are outdated as a training device changes or tasks change. No agency has formally assigned responsibility for keeping analyses current. Only the TPSC codes for the tank platoon have been updated by the Armor School when they published a revised MTP (ARTEP 17-237-10-MTP). Despite the aging problem and other shortcomings, analytic evaluations nevertheless continue to yield strong inferences about the extent to which missions and tasks will be trainable with devices and to what levels. Those kinds of inferences are prerequisites for specifying the sequences and mixes of device-based training and field training that form the core of Army

training strategies. These strategies define the role of training devices and simulators in accomplishing CATS objectives and set the stage for empirical evaluations and other experimentation to improve new Army training.

Analytic evaluations also can be used as-is. Comparing the instructional strengths and weaknesses of a device concept to the requirements document will yield recommendations for concept modifications. Such analyses can help to guide trade-off analyses in engineering design and can identify important training issues for operational testing. Later, the analyses yield recommendations for product improvements based on value added in terms of task coverage. The resulting priorities for device improvements, taken with associated costs of each recommended change, infuse objectivity into device proponents' decisions about which changes to buy. The decisions are subject to empirical validation later when the changes are tried out.

7. *Improve methods for documenting training by establishing one Army agency with adequate resources to perform the training analysis mission.*

The objectives of such an agency should be: (a) to collect and store diagnostic training data, (b) to perform routine and special analyses of training quality and effectiveness, and (c) to recommend methods, based on analysis results, of managing and conducting training to improve training processes and products.

The absence of an effective Army-wide system for archiving diagnostically useful training data is an obstacle to objectively estimating the relation between training resources and practices and the resultant effectiveness of the Army training system. Although massive databases on performance of units

at the CTCs have been archived, incomplete data have hampered productive analyses, resulting mainly from shortcomings of instrumentation and telemetry in live training. In addition, 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 research of Keesling, Ford, O'Mara, McFann, and Holz (1992) illustrates the value of such data: Their study suggested the effects of training resources available and training programs conducted at home station on performance of units at the National Training Center.

The fielding of computer-based training devices and the evolution of the Internet provides the infrastructure for a new start toward Army-wide monitoring and evaluation of training. Simulator sites with additional staff can serve as the focal point for collecting relevant local training data and administering surveys. These data and automated simulator data can be electronically uploaded to a central repository. More detailed data than were available to Keesling et al., will be needed to examine the role of the CCTT and other training devices in the CATS mix. Necessary data include documentation on training conducted at home station and field sites before and after units train at CCTT sites and at the CTCs. Records or data available on performance in training exercises also will be valuable, perhaps necessary, for evaluations that permit valid inferences about the effects of new training and its concomitants. Evaluators must establish solid working relations with units that are candidates for participation well in advance of installing the necessary data collection and storage systems.

The emphasis must be on obtaining and developing diagnostic information, not on collecting any and all data technology permits. One of the main jobs of our hypothetical analysis agency is to identify limited

sets of training process and product indicators for consistent monitoring over time. Such an agency also can design sophisticated sampling plans to obtain sufficient data while holding the costs of collecting data, storing databases, and performing analyses at an affordable level.

Establishing a central repository for home-station and other training data and for evaluation results would facilitate participation by Defense and Army agencies with analytic capabilities and responsibilities. To leverage its resources, the Army can establish mechanisms to authorize participation by qualified members of the commercial and academic communities with interests in investigating simulator and other training issues. Internet access to Army archives would encourage use of the data by interested and authorized users.

References

American Institutes for Research (1970). *Evaluative research: Strategies and methods.* Pittsburgh, PA: Author.

Babbitt, B. A., & Nystrom, C. O. (1989, June). *Questionnaire construction manual* (Research Product 89-20). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

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

Bessemer, D. W., & Myers, W. E. (1998). *Sustaining and improving structured simulation-based training* (Research Report 1722). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

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

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

Boldovici, J. A., & Bessemer, D. W. (1994, February). *Training research with distributed interactive simulation: Lessons learned from Simulation Networking* (Technical Report No. 1006). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

Boldovici, J. A., Harris, J. H., Osborn, W. C., & Heinecke, C. L. (1977, November). *Criticality and cluster analysis of tasks for the M48A5, M60A1, and M60A3 tanks* (Technical Report TR-77-A17). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences. (AD AO 48607)

Boldovici, J. A., & Kolasinski, E. M. (1997). How to make decisions about the effectiveness of device-based training: Elaborations on what everybody knows. *Military Psychology, 9*(2), 121-135.

Boldovici, J. A., Kraemer, R. E., & Lampton, D. R. (1986). *[Evaluation of a videodisk tank-gunnery trainer.]* Unpublished raw data.

Boldovici, J. A., Osborn, W. C., & Harris, J. H. (1977, Oct.). *Reliability in measuring unit performance.* San Antonio: Paper presented at US Military Testing Association.

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* (TRAC-WSMR-TEA-8-88). White Sands Missile Range, NM: US Army TRADOC Analysis Command.

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

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Boston: Houghton Mifflin Co.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational & Psychological Measurement, 20,* 37-46.

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

Cohn, V. (1994). Probable fact and probable junk. *NewsBackgrounder.* Los Angeles: Foundation for American Communications.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation : design & analysis issues for field settings.* Chicago, IL: Rand McNally College Pub. Co.

Cronbach, L. J. (1969). Evaluation for course Improvement. In R.C. Anderson, G.W. Faust, M.C. Roderick, D.J. Cunningham, and T. Andre (Eds.) *Current research on instruction.* Englewood Cliffs, NJ: Prentice-Hall.

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)

Efron, B. S. (1982). *The jackknife, the bootstrap, and other resampling plans.* Philadelphia: Society for Industrial and Applied Mathematics.

Embretson, S. E. & Hershberger, S. L. (Eds.) (1999). *The new rules of measurement: What every psychologist and educator should know.* Mahwah, NJ: Erlbaum.

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

Gagné, R. M. (1954). Training devices and simulators: Some research issues. *American Psychologist, 9,* 95-107.

Gagné, R. M., Foster, H., & Crowley, M. E. (1948). The measurement of transfer of training. *Psychological Bulletin, 45,* 97-107.

Gawande, A. (1998). The cancer-cluster myth. *The New Yorker,* 8 February.

General Accounting Office (1986, May). *Unit training: What it consists of, how it is evaluated, and how it is reported to the Congress* (GAO/NSIAD-86-94). Washington, DC: Author.

General Accounting Office (1990, September). *Test and evaluation: Improvements are being made in the Department of Defense's test planning.* (GAO/NSIAD-90-303). Washington, DC: Author.

General Accounting Office (1993, May). *Simulation training: Management framework improve, but challenges remain* (GAO/INSIAD 93-122). Washington, DC: Author.

Ghiselli, E. E. (1964). *Theory of psychological measurement.* New York: McGraw-Hill.

Gossman, J. R., Beebe, M. E., Bonnet, M., Forrest, D., Shadrick, S. B., Dannemiller, B., Mauzy, R. P., & Bonnet, M. (in publication). *The commander's integrated training tool for the Close Combat Tactical Trainer: Design, prototype development, and lessons learned.* Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

Grotte, J. H., Anderson, L. B., & Robinson, M. S. (1990, July). *Selected judgmental methods in defense analyses. Vol I: Main text* (AD E50-422). Institute for Defense Analysis: Alexandria, VA.

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

Harvey, E. (1999). Short-term and long-term effects of early parental employment on children of the National Longitudinal Survey of Youth. *Developmental Psychology, 35*(2), 445-459.

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

Holman, G. L. (1979). *Training effectiveness of the CH-47 flight simulator* (Research Report 1209). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

Horst, D. P., Tallmadge, G. K., & Wood, C. T. (1975). *A practical guide to measuring project impact on student achievement* (Contract No. DEC-073-6662). Washington, DC: US Department of Health, Education, and Welfare.

Houghton Mifflin Company (1984). *Webster's II New Riverside University Dictionary.* p. 976. Boston: The Riverside Publishing Company.

Hume, David (ca. 1760). An inquiry concerning human understanding.

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). Presido of Monetery, CA: PRC, Inc. and HumRRO, Inc.

Keppel, 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.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd Ed.). Monterey, CA: Brooks/Cole.

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: US 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.

Kuma, D., & McConville, L. (1982). *Independent evaluation report for M1/M60 series Unit Conduct of Fire Trainer (UCOFT)* (TRADOC ACN39373). Fort Knox, KY: US Army Armor Center.

Lawrence, D. H. (1954). The evaluation of training and transfer programs in terms of efficiency measures. *Journal of Psychology, 38,* 367-382.

Lindquist, E. F. (1953). *Design and analysis of experiments in psychology and education,* pp. 328-330. Boston: Houghton Mifflin.

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

Lopez, W. A. (1998, Spring). Rating scales and shared meaning. *Popular Measurement, 1*(1), 60.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley Publishing Co.

Luce, R. D. & Tukey, J. W. (1964). Simultaneous conjoint measurement. *Journal of Mathematical Psychology, 1,* 1-27.

Maxwell, S. E. (1994). Optimal allocation of assessment time in randomized pretest-posttest designs. *Psychological Bulletin, 92,* 778-785.

McCormick, E. J., & Bachus, J. A. (1952, April). Paired comparison ratings, I: The effect on ratings of reductions in the number of pairs. *Journal of Applied Psychology, 36,* 123-127.

Mitchell, R. (1979). *Less than words can say: The underground grammarian.* Boston: Little Brown.

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* (Research Product 88-03). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences. (AD A196 365)

Morrison, J. E., & Hoffman, R. G. (1992). *A user's introduction to determining cost-effective tradeoffs among tank gunnery training methods* (Research Note 92-29). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

Mudge, E. L. (1939). *Transfer of training in chemistry.* Johns Hopkins University Studies in Education, No. 6.

Murdoch, B. B. (1957). Transfer designs and formulas. *Psychological Bulletin, 54,* 313-326.

Myers, J. L. (1979). *Fundamentals of experimental design* (3rd Ed.). Boston: Allyn and Bacon.

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

Nunnally, J. C. (1967). *Psychometric Theory.* New York: McGraw-Hill.

Operational Test and Evaluation, Director (1998, Dec.). *Operational test and evaluation report on the Close Combat Tactical Trainer (CCTT).* Washington, DC: Author.

Orlansky, J. (1985). *The cost-effectiveness of military training.* Paper presented at the NATO Symposium on the Military Value and Cost-Effectiveness of Training, Brussels.

Patrick, J. (1992). *Training: Research and practice.* New York: Harcourt Brace Jovanovich.

PERT Program Evaluation Research Task Summary Report Phase I (1958, September). Washington, DC: Special Projects Office, Bureau of Naval Weapons, Department of the Navy.

PERT Program Evaluation Research Task Summary Report Phase II (1958, September). Washington, DC: Special Projects Office, Bureau of Naval Weapons, Department of the Navy.

Pfeiffer, M. G., & Horey, J. D. (1988). *Analytic approaches to forecasting and evaluating training effectiveness* (Tech. Rep. 88-027). Orlando, FL: Naval Training Systems Center.

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: US Army Research Institute for the Behavioral and Social Sciences.

Povenmire, H. K., & Roscoe, S. N. (1971). An evaluation of ground-based flight trainers in routine primary flight training. *Human Factors, 13*(2), 109-116.

Powers, T. R., McCluskey, M. R., Boycan, G. G., & Steinheiser, F., Jr. (1975). *Determination of the contribution of live firing to weapons proficiency* (FR-CDC-75-1). Alexandria, VA: Human Resources Research Organization.

Rambo, W. W. (1959). The effects of partial pairing on scale values derived from the method of paired comparisons. *Journal of Applied Psychology, 43,* 379-381.

Recer, P. (1999, 1 March). Working moms can rest easier. Orlando, FL: *The Orlando Sentinel.*

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

Roscoe, S. N. (1971). Incremental transfer effectiveness. *Human Factors, 13,* 561-567.

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

Saint Matthew. The gospel according to St. Matthew, 7:16. In National Publishing Company (1941), *The New Testament of Our Lord and Savior Jesus Christ* (p.16), Philadelphia, PA: Author.

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill.

Shavelson, R. J. (1988). *Statistical reasoning for the behavioral sciences.* Boston: Allyn and Bacon.

Sherikon Corporation (1995, May). *Task performance support (TPS) codes: A means to allocate MTP tasks to simulation-based collective training events* (Draft). Orlando, FL: Author.

Simpson, H. K. (2000). *Evaluating large-scale training simulations, Vol. 1,* (DMDC TR 00-05). Seaside, CA: Defense Manpower Data Center.

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 No. 904). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences. (AD A230 076)

Strunk Jr., W., & White, E.B. (1979). *The elements of style* (3rd ed.). New York: Macmillan.

Thompson, B. (1997). If statistical significance tests are broken/misused, what practices should supplement or replace them? Paper presented at the *105th Annual Meeting of the American Psychological Association.* Chicago, IL: American Psychological Association. (ERIC Document Reproduction Service No. ED 413 342).

Turnage, J., & Bliss, 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: US Army Research Institute for the Behavioral and Social Sciences. (AD A231 156)

Tversky, A., & Kahneman (1971). Belief in the law of small numbers. *Psychological Bulletin, 76*(2), 105-110.

US Army Training & Doctrine Command (1995, September). *Training development management, processes, and products* (TRADOC Regulation 350-70). Fort Monroe, VA: Author.

Venter, A. & Maxwell, S. E. (1999). Maximizing power in randomized designs when N is small. In R.H. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 31-58). Thousand Oaks, CA: Sage.

Wickham, J. A. (1983). Go the extra mile. *Soldiers, 38*(10), 6-10.

Wilkinson, L. & Task Force on Scientific Inference, APA Board of Scientific Affairs (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594-604.

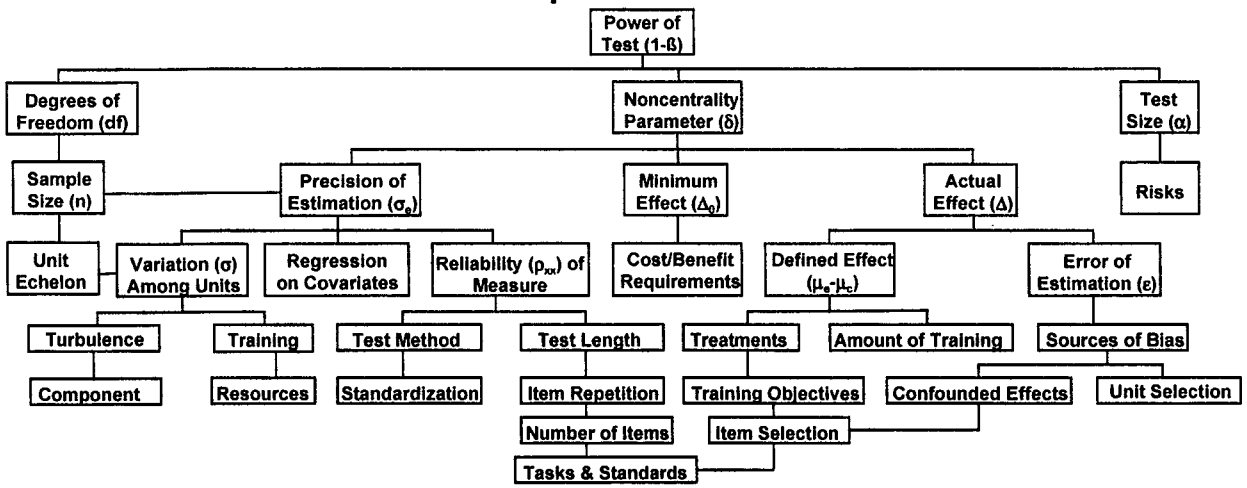
Winer, B. J., Michels, K. M., & Brown, D. R. (1991) *Statistical principles in experimental design* (3rd Ed.). New York: McGraw-Hill

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

Appendixes

APPENDIX A

The determinants of statistical power



APPENDIX B

Summary of Cook & Campbell's (1979) Methods for Reducing Within-Group Variance

If ways exist to deliberately manipulate and control the variable in question for every tested unit, then the variable can be equated at some constant value or condition for all units, thus preventing variation. For example, the time interval between the end of training and the beginning of testing easily is held constant by controlling the schedule of events in the field trial. The only shortcoming of this approach is that the observed results may be peculiar to the single time interval used in the test. If the treatment effect varies as a function of (i.e., interacts with) the controlled variable, then the obtained effect will not be representative of other treatment effects that would be obtained with other variable values. The evaluator should, therefore, have a strong rationale for wanting to estimate the treatment effect only under the single constant condition.

The second method for reducing within-group variance overcomes the limitation of the first method by deliberately using two or more values of the variable as a control factor in the test design. With each level of the control factor, treatment groups appear with independent samples of units. This design permits direct comparison of treatment groups at each level of the control factor, thus enabling examination of interaction effects and of the generality of the treatment effect. This method provides additional information on conditions associated with larger or smaller treatment effects. If the control factor does not cause variations in the treatment effect, this confirms the generality of the treatment effect with respect to the control factor. This method also increases statistical power by

removing variation caused by the control factor from the within-group variance.

In tests of new simulator training such as the CCTT, systematically varying the amount of training is essential. Doing so will help determine how much training is enough to produce a large performance benefit. If the amount of training is held constant at a value that is too small, the test results might wrongly suggest that the simulator training has little or no effect on unit performance.

Lacking any means to manipulate a variable, evaluators may use a third approach: Measure or describe the variables or conditions that occur spontaneously with each unit in the sample. One such variable might be a performance measure obtained in a recent field exercise that all units eligible for the planned test conduct on a regular basis. Such a measure partially reflects current unit proficiency. After measuring the variable before the treatment conditions are applied, selected sampling units make up sets with matching values, or subgroups stratified with similar values within levels. Then we can form equated treatment groups by placing in each treatment one unit from each matched set, or the same number from each subgroup level. The data analysis treats the matched sets or levels as a control factor, removing performance variation produced by the measured variable from the within-treatment variation.

The number of sampling units and the variable values that occur naturally may not be sufficient to form many well-matched sets or to form homogeneous stratified subgroups. Rather than discarding unmatched units from the sample, we can use regression methods (analysis of covariance, i.e., ANCOVA) with values for all units to adjust the results statistically; doing so removes performance

variations associated with the measured variable.[1] Evaluators also can use this fourth approach to adjust for environmental variables, unit characteristics, or measurement conditions that are measured concurrently with the training or testing. Such adjustment, however, is only valid if the evaluator is certain the treatment conditions cannot influence the measured variable. If, for example, one expects training treatments to cause variations in visibility or other weather conditions that occur during post-training testing, then adjusting for variations in test difficulty caused by weather variations is legitimate.

[1]More on the utility of ANCOVA is in Appendix E.

APPENDIX C

Scratch-Pad Estimates of Reliability and Validity

Background

The Sherikon Corporation (1995) developed a system of simulator ratings called the Task Performance Support Codes (TPSC). The TPSC estimate the capability of the Army's Close Combat Tactical Trainer to support units' performance of maneuver and other collective tasks for seven Battle Operating Systems (BOS). Sherikon based the TPSC rating system on the method Burnside (1990) used with SIMNET. Raters indicated their impression of the CCTT's ability to allow practice to criteria in subtask performance measures. The performance-measure ratings combine according to a set of decision rules to yield task-step ratings, which in turn combine according to a second set of decision rules, that is, the TPSC, to yield summary task ratings that range from 0 through 4. A summary 0 rating indicates the task in question is not at all supported by practice with CCTT. A rating of 4 indicates fully supported, that is, able to practice to MTP performance standards.

To demonstrate the use of TPSC, two retired colonels[1] independently made ratings that combined, as outlined above, to yield summary TPSC ratings for 64 tank platoon tasks. Table C-1 presents a summary of the two colonels' task ratings. The observed task counts entered in the main diagonal where both colonels assigned the

[1]One colonel was a former TRADOC System Manager for Combined Arms Tactical Training (TSM CATT); the other colonel was a team leader on Sherikon's contract for the Project Manager (PM) for CCTT.

same rating show that inter-rater agreement was good.

Table C-1.
Observed Agreement between the Two Colonels' Sets of Ratings Compared to Expected Chance Agreement

		Second Rater					
First Rater		0	1	2	3	4	Totals
4	Obs.	0	1	0	1	7	9
	Exp.	2.3	1.4	1.3	2.5	1.5	
3	Obs	0	0	3	13	3	19
	Exp.	4.8	3.0	2.7	5.3	3.3	
2	Obs	0	2	5	2	1	10
	Exp.	2.5	1.6	1.4	2.8	1.7	
1	Obs	4	2	0	2	0	8
	Exp.	2.0	1.3	1.1	2.3	1.4	
0	Obs	12	5	1	0	0	18
	Exp.	4.5	2.8	2.5	5.1	3.1	
Totals		16	10	9	18	11	64

	Difference				
	±0	±1	±2	±3	±4
Chance Agreement (%):	14.0	18.8	13.3	12.7	5.4
Actual Agreement (%):	61	31	6	2	0

Sherikon's Analyses

A discussion of inter-rater agreement is in the Sherikon report under the rubric, "Repeatability Analysis." This discussion includes the following statement (p. 19): "Results . . . indicate a high

correlation[2] between the two separate assessments. 61 percent of the resulting TPS Codes were the same and another 30 percent were +/- 1."

The percentage agreement scores are suggestive of high reliability, normally measured by a correlation coefficient. But to see the magnitude and judge the value of the obtained agreement and its associated reliability, we need to compute an index of agreement and a correlation coefficient and explore two characteristics not addressed in the Sherikon report. The first characteristic is the extent to which the obtained ratings differ from chance expectation; that is, the extent of agreement and reliability that could have resulted from using two random-number generators instead of two colonels. We can estimate the deviation from chance results by computing an agreement index and a correlation coefficient, and then computing the probability of having those values generated by chance. Doing so, however, does not immediately convey the concept of a chance distribution of ratings or the extent to which the agreement between the colonels' ratings differed from chance. Before computing the correlation coefficient, therefore, we shall do some arithmetic that we hope clarifies the underlying logic.

The second characteristic of reliability not addressed in the Sherikon report is the implication of reliability for estimating validity. We discuss this implication in a later section of this appendix entitled "Validity."

The Colonels' Ratings Compared to Chance

We may generate many different distributions of task ratings by randomly rearranging the counts in each row and column of Table C-1 while keeping the row and column totals unchanged. All such

[2]The Sherikon report did not present a correlation coefficient.

rearrangements constitute the entire set (population) of possible ratings conditioned on the observed totals. The observed ratings in the rows of Table C-1 (labeled "Obs.") are just one of these possibilities sampled by the colonels' ratings. For each cell in the table, the counts averaged over all possible outcomes in the population are the expected counts. The expected proportion for each cell can be calculated by multiplying the proportion of counts based on the corresponding column and row totals. Multiplying a cell's expected proportion by the total of the counts in the whole table then gives the expected count for that cell. The expected counts in the rows of Table C-1 (labeled "Exp.") show the average rating distribution resulting from chance.

The bottom two rows of Table C-1 compare the chance level of agreement and the actual agreement between the two colonels' ratings. These values are noteworthy in two respects: (1) the colonels' percentage of counts for rating differences of ±0 and ±1 far exceeds chance expectancy, (2) the colonels' percentages for the larger differences in ratings of ±2, ±3, and ±4 are much less than chance expectancy (a desirable outcome). Without question, the colonels' ratings show far better agreement than chance.

Cohen's K. One way of summarizing the overall degree of agreement between pairs of raters is to compute the widely used Kappa (K) index introduced by Cohen (1960). This statistic uses only the frequency of identical ratings; it applies to judgment dichotomies (yes/no or go/no-go) as well as rating scales. The idea of the index is to measure the increase in agreement above that expected by chance, relative to the maximum possible increase. With complete agreement $K = 1$. Chance is $K = 0$.

The formula for Kappa is:

$$K = \frac{n_o - n_e}{N - n_e}$$

In this formula, n_o is the observed total number of perfect agreements, n_e is the total expected number for the same cells, and N is the total number of observations for all cells. For the colonels' ratings data K = (39-14) / (64-14) = .50. The colonels agreed on 25 tasks more than the expected 14, but just half of the maximum of 50 tasks. The term "moderate" may therefore best describe the extent of the colonel's agreement. The estimated probability of a K value this large if the true $K = 0$ is $p < .001$, thus indicating that the index is statistically significant and confirming our impression that the colonels' agreement exceeded chance expectancy. A 99% confidence interval for K extends from .304 to .696.

Inter-Rater Reliability

As for replicability or repeatability – known as inter-rater reliability in statistical circles – the correlation coefficient used to quantify strength of relationship will account for how close rating judgments are when they disagree and for the prevalence of perfect agreement. Unlike the agreement index, a correlation coefficient credits the large percentage of the colonels' ratings that had differences of only ±1, compared to the small percentage that differed by ±2, ±3, or ±4. In the classic linear theory of test measures, any score (X) is assumed to be made up of a "true" score (t) disturbed by an added error component (e), so $X = t + e$. A reliability coefficient (r_{XX}) is a correlation coefficient computed from parallel or repeated measures obtained by the same method for the same set of things. A reliability coefficient is interpreted as a proportion, representing the ratio of the true score variance (σ_t^2)

to the error variance (σ_e^2). Similarly, the correlation between two judges' ratings provides a coefficient of inter-rater reliability that indicates the proportion of rating variance shared by, or common to, the two sets of ratings.

Spearman's r. By design, all correlation coefficients indicate the strength of the relationship or degree of correspondence between two properties measured for the same group of objects. The most common standard coefficient defined in statistics texts is the Pearson product-moment correlation, denoted by *r*. Statisticians have devised various other coefficients for specific situations or purposes: Spearman's coefficient, symbolized by r_S,, is a special case of Pearson's to be used when the properties are measured in the form of ordinal rankings, or when measures are converted to rank order numbers. The r_S coefficient is appropriate to show the strength of relationship between two sets of data on ordinal scales, that is, scales where a given numerical difference may not be exactly the same in all regions. For example, we may be confident that a difference of 4 is larger than a difference of 2, but have little confidence that 4 is precisely twice as large as 2. Scales such as those used by the colonels are ordinal.

Table C-2 shows the set of task ranks for each rater corresponding to the order of their ratings. Because the multiple task ratings grouped at each rating value have some tied order (inevitable with 64 tasks and only 5 possible ratings to choose from), the ranks assigned to each group are tied values. These rank values are the average of the ranks that fall in that group. Spearman's r_S quantifies how close the colonel's task rankings come to each other. The closer they are, the smaller the error components in the measures. An r_S of +1.0 indicates perfect agreement between the rankings for each task. An r_S of -1.0 indicates complete disagreement. A relation

of no positive or negative agreement results in $r_S = 0.0$, indicating that the rankings are randomly paired.

The formula for Spearman's correlation coefficient is:

$$r_S = 1 - \frac{6 \Sigma D^2}{N^3 - N}$$

Table C-2.
Two Colonels' Sets of Ratings Converted to Ranks

	Rating				
	0	1	2	3	4
First Rater					
Count	16	10	9	18	11
Rank	56.5	43.5	34	20.5	6
Second Rater					
Count	18	8	10	19	9
Rank	55.5	42.5	33.5	19	5

In the r_S formula given above, ΣD^2 is the sum of the squared differences between each pair of rankings, and N is the total number of objects ranked. This basic formula assumes no tied ranks, but its value may be corrected for the effect of ties by applying three additional formulas, which need not be elaborated here.[3]

[3]The formulas and rationales for their use are in Siegel's *Nonparametric Statistics for the Behavioral Sciences* (McGraw Hill, 1956, pp. 208-209).

Using the formula with the adjustment for tied ranks results in r_S = .83,[4] providing an estimate that error contributed only 17% of the total variation in ranks. This coefficient indicates good reliability, considering the fact that professional developers of commercial psychological tests usually regard r = .80 as acceptable (and, incidentally, a number that our experience suggests is greater than the reliabilities of scores from many field trials).[5] Thus by applying a few simple statistical procedures, we have managed to do a fair job of surrounding the issue of inter-rater reliability; the reliability of the colonels' ratings could be better, but not much better.

The utility of computing inter-rater reliability is, as mentioned earlier: The reliability coefficient is the hook we need to examine (1) statistical significance and (2) the potential validity of the ratings and of inferences therefrom.

Statistical Significance. Recall that in our earlier discussion of the colonels' ratings in Table C-1, we concluded agreement exceeded chance expectancy, and a statistically significant agreement index supported this conclusion. We have up to now obtained a reliability coefficient that is sizable. But we have not yet established whether our coefficient of agreement is statistically significant. The question (the null hypothesis) is whether the coefficient of agreement (inter-rater reliability) is unlikely if the rankings are actually randomly associated in the

[4]Comparing our earlier-reported K = .50 to r_S = .83, John E. Morrison (personal communication, November, 2000) noted, "I think the difference is because the former measures absolute agreement whereas the latter examines relative agreement. In other words, the two raters were evidently tapping into a common dimension, but they may have some differences with respect to precise calibration." We agree.

[5]Compare, for example, Powers et al.'s (1976) reporting the scores from their live-fire tank gunnery tests to be no better than "random guessing."

population so that the true value of the coefficient is zero. To make that determination requires calculating a *t*-statistic that has an approximate *t* distribution if the null hypothesis is correct. The *t*-statistic is computed using the following formula:

$$t_{N-2} = \frac{N-2}{r_S\sqrt{1-r_S^2}}$$

In the formula, the symbols retain their earlier definitions. The subscript with the *t* refers to the degrees of freedom (*df*), a parameter of the *t*-distribution used to estimate the probability of the statistic falling in certain value ranges. Use of the *t*-test to analyze the two colonels' ratings of the platoon tasks yielded $t_{62} = 11.55$. A value that large or larger could be expected by chance with probability less than .001.

We have now established the inter-rater reliability and the statistical significance of the ratings in question. The reliability is high but less than perfect, and is unlikely to have happened by chance, that is, the reliability is statistically significant. What may we legitimately conclude about the validity of the colonels' ratings?

Validity. There are several kinds of validity. For present purposes, let's use the definition commonly applied to concurrent or predictive validity: the correlation between a set of predictor scores and a set of scores from a more ultimate criterion. In our case, the predictor scores are the ratings from one or the other of the two colonels. The more-ultimate-criterion scores might be transfer measures from field trials that estimate the trainability of MTP tasks using CCTT. The strength of the correlation between predictor scores and more-ultimate criterion scores defines the validity coefficient. More-ultimate criterion scores, for example, from field trials,

obviously are not forthcoming in our case. We must therefore use indirect methods for estimating validity rather than directly calculating validity coefficients.

One indirect method involves estimating the theoretical or statistical limits on validity that would apply given our computed reliability of the two colonels' ratings. Implementing this method proceeds along the following lines: We have a set of predictor scores, that is, the two colonels' ratings. We wish we had criterion scores (e.g., from field demonstrations of trainability) that we could correlate with the colonels' ratings to yield a validity coefficient. But that wish is unlikely to come true, so we cannot compute the validity coefficient we would like to have. But we can compute (and have computed) the inter-rater reliability of our predictor scores ($r_{pp} \approx .83$).[6] And there is nothing to stop us from making assumptions, that is, from playing what-if games, about possible reliabilities of criterion scores (r_{cc}) if such scores were in fact available. Once we make those assumptions, it is easy to compute the maximum validity coefficients (r_{pc}) that could be achieved under each of our assumptions about the value of r_{cc}.

Examples of the What-If (Max-Validity) Game. First let's estimate the absolute maximum validity coefficient that could occur given any computed value of r_{pp}. To do so, simply assume our hypothetical criterion is perfectly reliable, that is,

[6]We have changed the notation from r_s in the previous section because we need a way to distinguish among predictor reliability (r_{pp}), criterion reliability (r_{cc}), and a validity coefficient (r_{pc}). The double subscripts are to remind us that r is always the correlation between two sets of scores, that is, two sets of predictors for r_{pp}, two sets of criterion scores (e.g., two halves of the same test or two separate tests) for r_{cc}, and one predictor reliability against one criterion reliability in the case of r_{pc}.

$r_{cc} = 1.0$. Then we apply Ghiselli's (1964, p. 271) Equation 9-14.

$$r_{pc} = \sqrt{r_{pp}\, r_{cc}}$$

Doing so demonstrates that the validity coefficient, that is, the correlation between a set of predictor scores with $r_{pp} = x$, and a set of criterion scores with $r_{cc} = 1.0$ cannot exceed $\sqrt{x}$. In our case, x is the inter-rater reliability of the colonels' ratings, approximately .83. The maximum possible validity of the colonels' ratings is therefore $r_{pc} = 0.91$.

Additional applications of Ghiselli's Equation 9-14 yield Ghiselli's Table 9.1 (1964, p. 271), reproduced here as Table C-3. Here we see again that for any two variables with $r_{pp} = x$ and $r_{cc} = 1.0$, the maximum correlation coefficient (validity coefficient in our case) is approximately $\sqrt{x}$. Thus with $x = .10$, .20, .30, and .90, the maximum values of the correlation coefficient with another variable having $r_{cc} = 1.0$ are .32, .45, .55, and .95, respectively.

Ghiselli's table obviously can be used to estimate maximum validity coefficients for values of r_{cc} other than 1.0. Assume, for example, we have calculated inter-rater reliability for some new TPSC ratings and found $r_{pp} \approx .70$. Assume also that we know that a field trial similar to one we might use to validate the TPS-code ratings yielded scores with r_{cc} identical to the reliability of our predictor, that is, $r_{cc} = .70$. Ghiselli's table shows that for the values $r_{pp} = .70$ and $r_{cc} = .70$, the maximum validity coefficient, r_{pc}, is .70.

Table C-3.
Ghiselli's (1964) Table 9-1. The Maximum Value of the Coefficient of Correlation (r_{pc}) between Two Variables in Relation to Their Reliability Coefficients

	r_{cc}										
r_{pp}	.00	.10	.20	.30	.40	.50	.60	.70	.80	.90	1.00
.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
.10	.00	.10	.14	.17	.20	.22	.24	.26	.28	.30	.32
.20	.00	.14	.20	.24	.28	.32	.35	.37	.40	.42	.45
.30	.00	.17	.24	.30	.35	.39	.42	.46	.49	.52	.55
.40	.00	.20	.28	.35	.40	.45	.49	.53	.57	.60	.63
.50	.00	.22	.32	.39	.45	.50	.55	.59	.63	.67	.71
.60	.00	.24	.35	.42	.49	.55	.60	.65	.69	.73	.77
.70	.00	.26	.37	.46	.53	.59	.65	.70	.75	.79	.84
.80	.00	.28	.40	.49	.57	.63	.69	.75	.80	.85	.89
.90	.00	.30	.42	.52	.60	.67	.73	.79	.85	.90	.95
1.00	.00	.32	.45	.55	.63	.71	.77	.84	.89	.95	1.00

Note: From *Theory of Psychological Measurement* (p. 271) by E.E. Ghiselli, 1964, New York, McGraw-Hill Book Company. Copyright 1964 by McGraw-Hill Book Company. Reprinted by permission.

As a final what-if exercise, let's assume evaluators are, as we believe they should be, sufficiently immersed in the results of Army field trials to know the average reliability of field-trial or other criterion scores, that is, r_{cc}. For this example, let $r_{cc} \approx .65$.[7] Ghiselli's table shows that for our observed $r_{pp} = .70$ and with field-trial r_{cc} estimated at .65, the maximum validity coefficient is .68. It is thus possible to estimate maximum validity coefficients by making reasonable assumptions about r_{cc} even when no criterion measures are available. This possibility has

[7]Performing analyses such as suggested here would be easy if evaluators would routinely report the reliabilities of compared groups' scores in their field trials or would report complete sets of raw scores so interested readers could compute reliabilities.

obvious implications for estimating the maximum validities of SME ratings, of model outputs based on SME estimates, and of the results of field trials: All that is needed is (a) to compute the reliability of our observed predictor scores (b) make tenable estimates about the reliabilities of criterion scores, and (c) estimate maximum validity from Ghiselli's table.[8]

It is important to emphasize that the values in Ghiselli's table are upper limits; even with $r_{pp} = 1.0$, the validity coefficient may be zero if $r_{cc} = .00$. The implications for estimating validity coefficients from predictor reliability, r_{pp}, are (a) high r_{pp} does not guarantee a high validity coefficient, (b) low r_{pp} guarantees low validity, and (c) low r_{pp} can lead to a high estimate of maximum validity only if we make unrealistic assumptions about r_{cc}, for example, $r_{cc} \geq .90$.

[8]We strongly recommend applying similar procedures to estimate the maximum validity of scores from field trials.

APPENDIX D

Repeated-Measure Latin Squares

A special kind of Latin square is the repeated-measure Latin square. This design takes its name from the repeated measurement of each sampling element's performance on several successive occasions. Individual sampling elements or randomly selected groups of sampling elements form the rows of the square, while the successive occasions form the column variable of the square. The individual or group in each row receives each treatment condition on one occasion. Every row gets a different sequence of treatments, counterbalanced so that every treatment appears exactly once in each column and once in each row of the square. This design controls variation among sampling elements as well as differences or trends in performance that result from changes in sampling units or conditions that differ across occasions. Such situations are common in military training, inasmuch as units' performance may change because of learning or forgetting, or some other external conditions may differ consistently among occasions, such as time of day.

Table D-1 illustrates a repeated-measure Latin square design for an experiment in the CCTT intended to assess the ability of armored cavalry units to perform a movement to contact mission against hypothetical future enemy units. The rows include six cavalry troops from two squadrons, one troop assigned to each sequence of treatments. The columns are the orders of exercises completed on six consecutive occasions across three days. Exercises with odd orders are done in the morning and even orders in the afternoon. Repeated measures of performance use the same dependent variable (e.g., loss-exchange ratio) for each exercise. The treatments are six

different scenarios that combine three radically different kinds of OPFOR organizations and tactics with two operational environments: desert or temperate terrain.

Table D-1.
Example of a Repeated-Measure Latin Square Design

	Exercise Order					
Unit (Sequence)	1st	2nd	3rd	4th	5th	6th
Sqn 1, Trp A (I)	A	B	C	D	E	F
Sqn 1, Trp B (II)	B	C	A	F	D	E
Sqn 1, Trp C (III)	C	A	B	E	F	D
Sqn 2, Trp A (IV)	D	F	E	B	A	C
Sqn 2, Trp B (V)	E	D	F	A	C	B
Sqn 2, Trp C (VI)	F	E	D	C	B	A

Note. Letters in each cell of the square refer to exercise scenarios (treatments) with varied OPFOR and terrain.

Repeated-measure Latin squares have several advantages over other kinds of designs for training and transfer evaluations. First, any number of treatments fit into a repeated-measure Latin square because more occasions allow any size square. The practical limitations are the number of sampling units available and the time required for each unit to complete the treatments. The repeated-measure Latin square thus avoids the limitations of other designs on the number of sampling elements imposed by the structure of units in echelons.

Second, repeated-measure Latin square designs tend to be statistically powerful, because they completely control the variations among sampling elements produced by all the unique characteristics of particular elements. The random variations that remain to influence the difference among treatments are often a small fraction of the variation in a

completely randomized or in a randomized block design.

Third, repeated-measure Latin square designs provide a convenient way to manipulate amount of training jointly with other treatments in evaluations of training effects, including transfer of training. In the design shown in Table D-1, a progressive increase in performance across columns shows the average effect of the amounts of training that accumulate as additional scenarios are completed. With performance measures from simulator exercises, increasing performance suggests within-device transfer.

Both ordinary and repeated-measure Latin square designs can accommodate additional independent variables by using additional squares. Additional squares may control nuisance variables or permit examination of treatment effects of interest to evaluation proponents. An evaluation of three kinds of AAR equipment and procedures, for example, might use one square as in Table D-1 with each of the AAR treatments. Differences in the performance trend across columns would indicate a training effect of AAR treatments. Differences among groups of units assigned to different squares in performing a subsequent field exercise would show effects on cumulative transfer of training.

On the negative side, Latin square designs require special assumptions for valid estimates of effects and tests of significance. Latin square designs counterbalance only 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 include, for example, 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 the validity of specific assumptions about the nature of the variables and their interactions. Expert advice is imperative when planning, analyzing data, and interpreting the results.

An additional complication in repeated measure Latin square designs is that multiple measures taken from the same sampling elements are not independent. Special methods of statistical analysis may be necessary depending on the pattern of correlations among the columns. In addition, prior treatments in a sequence may alter the effect of a later treatment. 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 methods or peculiar design arrangements may be required to prevent or correct for carry-over effects.

APPENDIX E

Transfer-Efficiency and Savings Estimates

Never use correlation between training scores and test scores to estimate transfer. As Mudge noted in 1939 and Gagné iterated in 1954, correlations do not establish the causal link necessary for demonstrating transfer. A high positive correlation between training scores and test scores only suggests that Ss used similar skills in training and in testing, not that training caused the test scores.

In addition to misusing correlation, some training researchers and evaluators mislead readers with transfer formulas. The chief abuses are failure to report conventional analyses of raw scores in addition to the results of transfer formulas and ignoring various deficiencies of transfer formulas that lead to "spurious quantitative reasoning" (Gagné, Foster, & Crowley, 1948, p. 98).

Transfer formulas yield estimates of percentage transfer, which is computed using a T (for transfer) group mean score or an E (experimental) group mean score and a C (control or conventional) group mean score. The means are variously combined in one or more of four classes of formulas:

(a) The first class of formulas references the T mean against the C mean. An example of formulas in this class, for cases in which a higher score means better performance than a lower score, is:

$$\text{Percentage Transfer} = ([T - C] / C)(100).$$

The formula yields an asymmetric distribution of scores, ranging from plus infinity to minus 100; it is

therefore more suitable for negative transfer than for positive transfer.

(b) The second class of formulas references the T mean to the maximum possible gain. An example for cases in which high scores mean better performance than low scores is:

$$\text{Percentage Transfer} = ([C - T] / C - \text{Max})(100).$$

The distribution of scores is once again asymmetric, but this time ranges from plus 100 to minus infinity.

(c) The third kind of formula was developed by Murdoch (1957), who wanted a formula that would yield symmetrical distributions of positive and negative scores with identical absolute values of upper and lower limits, "preferably, of course, 100%" (p. 322). Murdoch's solution, when higher scores mean better performance, was:

$$\text{Percentage Transfer} = ([T - C] / T + C)(100).$$

The three classes of formulas discussed so far yield estimates of relative amounts of transfer. (Absolute measures of transfer do not exist and are not forthcoming.) Expanding or otherwise changing the formulas to reflect the cost (in number of trials, amount of training time, or dollars) of the T and C groups' reaching given levels of proficiency is a simple matter. Depending on whether we divide percentage transfer into or by dollars, for example, we can estimate price per percentage transfer or percentage transfer per dollar, both of which exemplify the fourth class of formulas: efficiency or cost-effectiveness. Lawrence described various transfer-efficiency formulas in 1954. The transfer-efficiency formulas with which most are familiar are the transfer-effectiveness ratio (TER) introduced by

Povenmire and Roscoe (1971), with later variations by Roscoe, also in 1971.

(d) The TER is used to estimate the savings, in trials or time (translatable to dollars), to be realized by using a training device before learning to use parent equipment such as aircraft. The TER is computed by dividing the mean number of simulator trials or amount of time for the T group into the difference between the mean numbers of aircraft trials or amount of time for the C and T groups; thus:

TER = (WS[1] trials for C)-(WS trials for T)/
(Sim trials for T).

A TER of 1.0 indicates, according to its inventors, the device is as efficient as the aircraft for training. A TER greater than 1.0 indicates the device is more efficient than the aircraft or other parent equipment. And a TER less than 1.0 indicates the training device is less efficient than the parent equipment.

One of the presumed benefits of savings measures, efficiency estimates, cost-effectiveness formulas, and various other combination scores is that they yield bottom lines that are easy to read. Whenever we see one of these algorithms, we should ask, "What conditions must be fulfilled to yield identical scores?" Then plug in various numbers to yield identical scores and ask yourself whether the compared things, concepts, are in fact identical from a rational point of view. In the case of the TER, for example, the condition that must be fulfilled to yield a ratio of 1.0, that is, device and weapon system equally efficient, is that any combination of weapon-system and simulator trials or training time for T be the same as the mean number of weapon-system trials or training time for C. Let's make up some numbers to yield

[1]Weapons Systems

identical TERs of 1.0 and then examine the underlying logic to see if it makes sense:

(a) A C group requires 20 WS trials to reach criterion.

(b) A T group using Device A requires 18 simulator trials and 2 WS trials to reach the same criterion as the C group.

(c) A second T group using Device B requires 2 simulator trials and 18 WS trials to reach the same criterion as the C group and the other T Group.

All combinations of weapon-system and simulator trials total 20. Both devices are therefore as efficient as the weapon system. And both devices are, by definition, equally efficient--even though the number of weapon-system trials required after practice with Device B was 9 times the number required after practice with Device A. This makes no sense and accounts for our skepticism when it comes to cost-effectiveness evaluations (a field fraught with conceptual difficulties exemplified by the Kennedy-Johnson-era whiz kids' demonstrating that cost-effectiveness analyses could support any position you wanted them to).

Additional considerations that should make us question reflexively all reports of transfer-efficiency or savings estimates are:

(a) The numerator in each of the four kinds of formulas discussed above is the mean difference between the experimental and control groups' scores. (That seems true for all transfer formulas, but we're not sure.) Researchers who use transfer formulas hardly ever report the reliability of the difference scores used in their numerators. (See, for example, the 100 or so TERs for the Chinook simulator reported by Holman, 1979.) Our thinking is that if the difference

is unreliable, then the numerator should be zero, and a re-examination of research results reported in terms of transfer formulas would show that in many of the studies no reliable transfer was produced. How to interpret the results of transfer formulas is unclear, as is justification for using those results in extrapolations of savings and other simulator benefits.

(b) The difference scores used in the numerators of transfer formulas do not reflect where trainees began or where they ended.

(c) The difference scores used in the numerators compound the unreliability of the two scores from which they are derived. If evaluators do not report the reliabilities of their T and C scores, we have no basis for estimating the reliability of the difference score, or of its validity, or of the validity of inferences from a TER or other savings formulas.

(d) Comparisons of gains for device and conventionally trained groups provide no indication of how much, if any, of the gain was due to training; that is, the C groups are not control groups. They are, in fact, treatment groups. Establishing how much, if any, of the gain was due to training requires a no-training control group or, at least, using each group as its own control – as with pre- and post-intervention tests, for example. Think of this in terms of toothpaste research: We find no difference between the effects of brushing with Colgate and brushing with Crest – an uninteresting finding, compared to the important questions that remain unanswered – e.g., What about the effects of brushing with water? Or not brushing at all?

Dilettantes are fond of transfer and savings estimates, not only because the estimates yield bottom lines that are easy to read, but also because the arithmetic

looks like the computations of unit prices in supermarkets. The similarity is irrelevant, however, because we have only relative, and no absolute, measures of transfer. There is no glass-encased, universally accepted unit of transfer at the Bureau of Standards (or anywhere else), as there are universally accepted standards for what constitutes an ounce, a pound, and all other measures that make unit pricing a semi-rational basis for comparing identical products. But does anyone believe that 20% transfer to Weapon System A is in any way identical to 20% transfer to Weapon System B? Or to 20% transfer to Weapon System A measured at another time? Or half as good as 40% transfer? We hope not.

Our advice is to avoid using transfer formulas and transfer-efficiency formulas such as the TER. Use conventional analyses of raw scores instead. If whoever is paying for the evaluation insists on using transfer formulas, then (a) use the formulas as supplements to conventional analyses of raw scores, and (b) report, for naive readers' benefit, the limitations discussed above. Gagné et al. (1948), comparing the measures provided by transfer formulas with the raw scores from which they are derived, noted, "The utilization of raw score values to express transfer is a procedure which has a number of advantages, chief among which is precision of meaning" (p. 98). Precision of meaning is, in our view, what separates evaluation reporting that can be believed from evaluation reporting that cannot. Evaluators who use transfer and efficiency formulas are free to do conventional analyses of raw scores as well. When evaluators do not report conventional analyses of raw scores, we wonder, and hope you will too, whether the evaluators don't know any better or are trying to bamboozle us readers. Either way, the light shed by their work on transfer is not worth the candle.

APPENDIX F

ANCOVA: Additional Considerations, Utility for Evaluating Simulator Training

Additional Consideration 1: The Disordinal Interaction

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 an interaction between covariate and treatment effects, 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, that is, the slopes of the lines differ. Such interactions become important when the lines cross. Interactions with crossing lines are termed "disordinal" because the treatments' rank order determined by their dependent variable values differs at different levels of the covariate.

Figure F-1 illustrates this sort of disordinal interaction pattern, with the crossover occurring at a covariate value near 40. In Figure F-1, the line for Treatment A increases slowly, crossing under a more steeply rising line for Treatment B. If the treatments are alternative training conditions, this pattern indicates that Treatment A is better for units with pretest scores below 40, but that Treatment B is better for units with scores above 40. The obvious implication for training policy is that the diagnostic pretest should determine units' assignment to the most beneficial training condition.

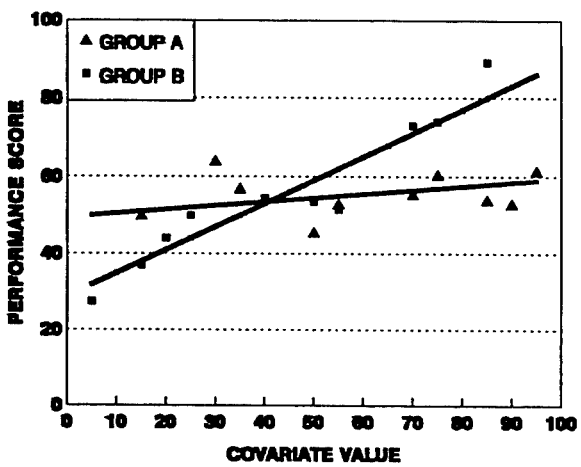


Figure F-1. Example of disordinal interaction between treatment groups and covariate in ANCOVA.

Additional Consideration 2: ANCOVA Cautions

The evaluator planning to use ANCOVA must observe three cautions. First, the evaluation design must insure that the independent variables manipulated in the evaluation cannot causally influence the values of the covariate. Measuring the covariate before administering the treatments will meet this condition. Otherwise, removing variation in the dependent variable associated with the covariate will also remove part of the treatment effect.

Second, the covariate cannot affect the nature of the independent variable 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 advertently, with good intentions, or 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 most experienced leaders. If an increase in training or test difficulty reduced the posttest scores for those platoons, the posttest difference between treatment groups would be smaller as platoon leader experience (the covariate) increased, leading to an incorrect conclusion. If such side effects of the covariate are difficult to prevent, the evaluator should at least seek to detect them. Whenever variations are possible, he should measure if possible, those aspects of the treatment and test conditions actually administered.

Third, randomly assign the sampling units to treatment groups. When intact groups of sampling units (e.g., companies when platoons are measured) form treatment groups, any effects produced by preexisting differences between the intact groups are entirely confounded with the effects of the independent variable (e.g., training). 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. Cook and Campbell (1979) fully treat the problem of drawing valid inferences from ANCOVA designs that use intact groups or nonequivalent control groups. Evaluation designs should avoid such confounded arrangements except when no other alternatives are possible.

Utility of ANCOVA for Evaluating CCTT

One kind of covariate that should be useful in tests and evaluations of unit training with simulators such as the CCTT is a pretest measuring performance on tasks one echelon below the organizational level of the units to be trained. If the units sampled in the treatment conditions are companies, for example, then the pretest should include platoon-level tasks related to the company missions trained in the simulator after the pretest. ANCOVAs then can use the pretest scores in analyses performed both on simulator–based training performance measures and on field performance measures.

The objectives of such analyses are twofold, serving both purposes discussed previously. The first objective is to determine if low levels of performance on the pretest tasks will prevent effective simulator training and transfer of training below some specific score value. These results help define what minimal proficiency at the lower echelon is needed to make simulator training worthwhile at the higher echelon. Such information forms the basis for efficient management of multiechelon training.

The second objective of using lower echelon pretest scores as a covariate is to determine whether there are disordinal interactions between pretest and treatment conditions. These results help determine how best to conduct training with units coming to the simulator with varying proficiency on lower echelon tasks. Such information can help maximize the benefits derived from a given investment in simulator training.

Appendix G

The Quasi-Experiment: Estimating Transfer of New Training (e.g., CCTT) to Units' Field Performance

The quasi-experiment described here is appropriate for the CCTT with armor and mechanized infantry units rotating to the National Training Center, Ft. Irwin, CA. The design also can be used at The Joint Readiness Training Center, Ft. Polk, LA, when CCTT enhancements fully support light infantry units. The first step toward creating the design in advance of fielding new training is to adopt a diagnostic pretest exercise and a standard data collection system using both observers and instrumentation. For the CCTT, for example, both armor and mechanized infantry company exercises would be developed to provide measures of essential performance elements (tasks, subtasks, or standards) for platoons and companies. An expert panel with representatives of the branch schools, PM-CATT, TSM-CATT, and the CTC would then meet and eventually agree that the performance elements are among those with the highest priority for CCTT training. The panel also must agree that the measures of performance are valid under the conditions presented in the CTC exercises.

In advance of fielding CCTT or other new training, each company rotating into the training is pretested before participating in other exercises planned for that rotation. Baseline data collection starts at least one year before the first newly trained units appear at the CTC. Evaluators obtain data on home-station training for these same units. These baseline data enable preliminary analyses of time trends for performance measures, including possible seasonal variations. Analyses of the home-station training data also identify correlates of performance,

including any consistent differences among units from different locations.

After fielding the new training to various locations, the units at those locations and trained in the CCTT will begin to appear at the CTC among units from other locations that had no CCTT training. Improvement in performance for the CCTT-trained units compared to the non-CCTT-trained units provides evidence of transfer of CCTT training to field performance in the CTC pretest exercise. At the same time, evaluators analyze records of CCTT use in relation to variations in unit performance to identify the kinds and amounts of CCTT training that produce various kinds and amounts of transfer.

The inference that the CCTT training is responsible for any transfer effect in our quasi-experiment may be invalid. Reservations arise because neither the CCTT-trained group of units nor the control group of units were randomly sampled, and neither group trained at randomly sampled times. Evaluators must therefore analyze CTC records to determine if systematic differences among training occasions in environmental variables or measurement conditions biased the group comparison. If seasonal or other trends appear in the baseline, the results will have to be adjusted for these effects. Evaluators must also compare the performance and home-station training data for the control group to the baseline data for units from the same location to show that the control units were similar to the baseline units before the quasi-experiment began.

Similar comparisons also must show the CCTT-trained group is similar to the baseline units in all ways except those related to CCTT training. In particular, the home-station training for the CCTT-trained group may have changed substantially from the baseline to accommodate the time required for the CCTT in the units' training schedules. Objective

evaluators will conclude only that the CCTT training in combination with associated changes in unit training produced the observed effects. Depending on the nature of the effects and the pattern of home-station correlates of performance, separating the contribution of individual factors may not be possible.[1]

The main costs of the evaluation in our example are devoted to gathering and analyzing home-station training data. Data collection at the CTC uses resources that are already in place. Commanders probably will find the pretests beneficial, because the scores will point to subordinate units' strengths and weaknesses and can help the commanders focus these units' training during the remaining CTC exercises. Despite the reservations that attend quasi-experiments, such evaluations are almost certain to provide valuable diagnostic information about beneficial unit training practices, in and out of the new training device.

[1]See earlier comments regarding interaction effects and the need for expert statisticians.

Appendix H

Analytic Methods: Three Examples

The examples of analytic methods summarized here were the work of Burnside (1990), of Drucker and Campshure (1990), and of Sherikon (1995). All presented tenable recommendations for immediate improvements in new Army training at prices considerably lower than prices that attend field trials.

Burnside (1990) and Drucker and Campshure (1990) analyzed the strengths and weaknesses of SIMNET for training specific tasks defined in ARTEP Mission Training Plans (MTP). 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 corresponding weapons systems operated in a field-training environment. Drucker and Campshure also made tenable educated guesses about the effects of the similarities and differences on transfer to field exercises.

Burnside (1990) developed and used a rule-based method to estimate which MTP standards could be met and which subtasks and tasks could be performed in SIMNET. Burnside's work rated a standard "highly supported" (H), for example, if the standard 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. Other rule-based categories included Partially (P), Minimally (M) and Not (N) supported, and Outside (O) support required. 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 tasks.

Sherikon (1995), at the request of PM-CATT, performed an analysis similar to Burnside's for the CCTT. Sherikon defined Task Performance Support Codes (TPSC) by mapping categories similar to Burnside's into a 5-point scale (0-4). Sherikon's experts produced ratings suggesting that CCTT would provide high or moderate support for 77% of battalion task force tasks,[1] 67% of company team tasks, and 54% of platoon tasks.[2] Additional description of Sherikon's work is in our Appendix B.

[1]The extent of CCTT support for battalion tasks may be misleading as an indication of support for unit training at this level because most sites will not have enough vehicle modules to deploy a full battalion. The ratings are accurate for tasks defined at battalion level, but some parts of the battalion task force must be filled in by MODSAF entities controlled from a workstation.

[2]The TPSC codes recently were incorporated as guidance in the Commander's Integrated Training Tool software designed to support selection, modification, and creation of CCTT exercises (Gossman et al., 1998).