

AD-A160 669

AN ANALYSIS OF EVALUATOR BIAS IN THE MARINE CORPS
COMBAT READINESS EVALUATION SYSTEM(U) NAVAL
POSTGRADUATE SCHOOL MONTEREY CA G M WHEELER ET AL.
AUG 85 NPS-54-85-016

1/1

UNCLASSIFIED

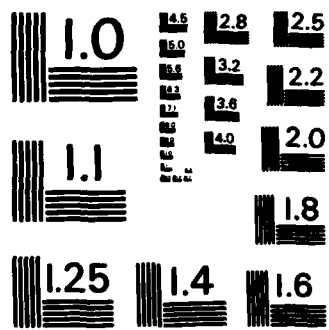
F/G 5/9

NL

END

FILMED

DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS - 1963 - A

NPS-54-85-016

NAVAL POSTGRADUATE SCHOOL

Monterey, California



AD-A160 669

DTIC FILE COPY

DTIC
ELECTE
OCT 29 1985
S
A

AN ANALYSIS OF EVALUATOR BIAS IN THE MARINE CORPS
COMBAT READINESS EVALUATION SYSTEM

by

George M. Wheeler
Joseph F. Mullane, Jr.
and
Kenneth J. Euske

Approved for public release; distribution unlimited

Prepared for: Commandant of the Marine Corps
Headquarters, U.S. Marine Corps
Washington, D. C. 20380

85 10 29 014

NAVAL POSTGRADUATE SCHOOL
Monterey, California

Commodore R. H. Shumaker
Superintendent

David A. Schradz
Provost

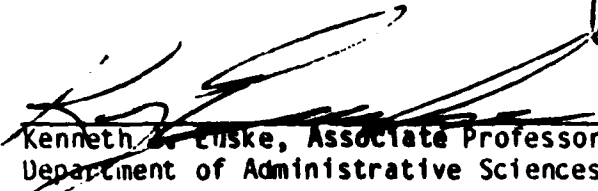
The research summarized herein was sponsored by the Commandant of the Marine Corps (Code PUR)

Reproduction of all or part of this report is authorized.

This report was prepared by:


George M. Wheeler, CAPT, USMC
Department of Administrative Sciences


Joseph F. Mullaney Jr., COL, USMC
Department of Administrative Sciences


Kenneth A. Encke, Associate Professor
Department of Administrative Sciences

Reviewed by:


Willis R. Greer, Chairman
Department of Administrative Sciences

Released by:


Kneale T. Marshall
Dean of Information and Policy Sciences

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

AD-A160669

REPORT DOCUMENTATION PAGE

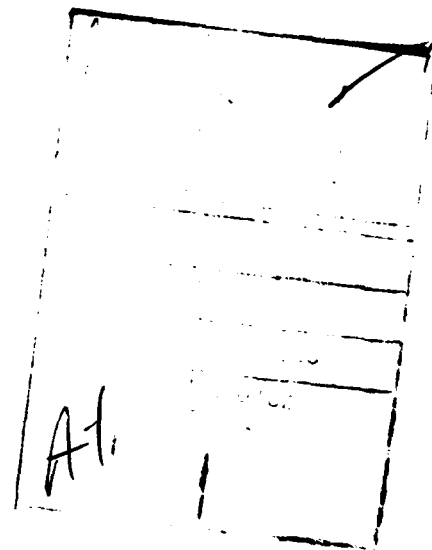
1a REPORT SECURITY CLASSIFICATION UNCLASSIFIED		1b RESTRICTIVE MARKINGS	
2a SECURITY CLASSIFICATION AUTHORITY		3 DISTRIBUTION/AVAILABILITY OF REPORT	
2b DECLASSIFICATION/DOWNGRADING SCHEDULE			
4 PERFORMING ORGANIZATION REPORT NUMBER(S) NPS-54-84-016		5 MONITORING ORGANIZATION REPORT NUMBER(S)	
6a NAME OF PERFORMING ORGANIZATION Naval Postgraduate School	6b OFFICE SYMBOL (if applicable)	7a NAME OF MONITORING ORGANIZATION	
6c ADDRESS (City, State and ZIP Code) Monterey, CA 93943		7b ADDRESS (City, State and ZIP Code)	
8a NAME OF FUNDING/SPONSORING ORGANIZATION HQ, U.S. Marine Corps	8b OFFICE SYMBOL (if applicable) POR	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER	
8c ADDRESS (City, State and ZIP Code) Washington, D.C. 20380		10 SOURCE OF FUNDING NUMBERS	
		PROGRAM ELEMENT NO	PROJECT NO
		TASK NO	WORK UNIT NO
11 TITLE (include Security Classification) AN ANALYSIS OF EVALUATION BIAS IN THE MARINE CORPS COMBAT READINESS EVALUATION SYSTEM			
12 PERSONAL AUTHOR(S) George M. Wheeler, Joseph F. Mullane, Jr. and Kenneth J. Euske			
13a TYPE OF REPORT Technical Report	13b TIME COVERED FROM 1984 TO 1985	14 DATE OF REPORT (Year, Month, Day) August 1985	15 PAGE COUNT 61
16 SUPPLEMENTARY NOTATION			
17 COMBAT CODES		18 SUBJECT TERMS (Continue on reverse if necessary and identify by block number)	
FIELD	GROUP	SUB-GROUP	
19 ABSTRACT (Continue on reverse if necessary and identify by block number) The Marine Corps Combat Readiness Evaluation System (MCCRES) was designed to provide timely and accurate information on Marine Corps operating units ability to carry out their assigned combat missions using 'expert' evaluators to observe and grade simulated combat operations. This study examines the MCCRES for bias susceptibility which would cause inaccurate evaluation by addressing two questions: (1) Can evaluator factors subject to bias be identified, and (2) how can these factors be controlled or accommodated?			
20 DISTRIBUTION/AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS		21 ABSTRACT SECURITY CLASSIFICATION	
22a NAME OF RESPONSIBLE INDIVIDUAL		22b TELEPHONE (include Area Code)	22c OFFICE SYMBOL

DD FORM 1473, 84 JAN

63 APR EDITION MAY BE USED UNTIL EXHAUSTED
ALL OTHER EDITIONS ARE OBSOLETEUNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

PREFACE

The research effort represented by this report was funded by the Commandant of the Marine Corps. The objective of the research effort was to evaluate feedback mechanisms for MCCRES. Hopefully, the reader will judge that the objective has been satisfied. In our opinion it has been exceeded due largely to the formal and informal support offered us by the Marine Corps officers.



AN ANALYSIS OF EVALUATOR BIAS
IN THE MARINE CORPS COMBAT
READINESS EVALUATION SYSTEM

by

George M. Wheeler
Captain
United States Marine Corps

Joseph F. Mullane
Colonel
United States Marine Corps

and

Kenneth J. Euske
Associate Professor of Accounting

Department of Administrative Sciences
Naval Postgraduate School
Monterey, CA 93943

EXECUTIVE SUMMARY

The Marine Corps Combat Readiness Evaluation System (MCCRES) was designed to provide timely and accurate information on the ability of Marine Corps operating units to carry out their assigned combat missions. This study examines the MCCRES for susceptibility to bias which could cause inaccurate evaluations.* Two primary questions were addressed: 1. Can factors that influence evaluator bias be identified? 2. How can these factors be controlled or accommodated?

As a point of reference, the investigators developed a working definition of evaluation, synthesizing several definitions. They define evaluation as: A judgment of some program with the purpose of contributing to decisions concerning the current attainment of that program's goals or objectives.

The concept of evaluation was reviewed using academic research. Principle, approaches, and training concepts were analyzed for information helpful or relevant to MCCRES' evaluations. It was found that while the overall purpose of diverse types of evaluations may be the same, (i.e., providing information to aid in decision making) different situations may call for different approaches to provide the necessary information. Techniques can be chosen to: fit evaluation to evaluator's skills (quasi-legal vs. professional review approaches); fit evaluation to program objectives (system analysis vs. behavioral-objectives approaches); or even to ignore evaluation goals (goal-free approach).

The subject of evaluators is then analyzed. Types of evaluator errors, evaluator sources, error sources and error reduction techniques are reviewed.

Errors are discussed in terms of variable errors and constant errors. Variable errors are differences in evaluation scores of specific items of an evaluation due to different evaluators or evaluations over time. Constant errors are classified as 'halo' errors (evaluation on the basis of overall impressions), as 'central tendency' errors (evaluators rate all near the middle grade), and 'leniency' errors (or its opposite, too strict error). Sources of evaluators (superiors, peers, and disinterested parties) tend to introduce bias unique to each group. Evaluation by superior may lead to direct reward or punishment. Peer evaluation appears to offer great benefits to an evaluation program despite the perception of a friendship bias. Disinterested parties, often facilitating a more objective evaluation, may have a limited insight into the factors which indicate good job performance. Techniques to reduce error (e.g., training and testing of evaluators and taking measures to reduce the subjectivity of evaluation measures) are reviewed.

The MCCRES is then examined to identify areas where errors or bias may be injected. The MCCRES is compared with the 'professional review' approach since it is desirable that evaluators have recently served successfully in billets relating to the functions they are to observe. Three main roles of evaluators are: exercise controller; umpires; and performance evaluators. Task performance is evaluated in terms of Mission Performance Standards (MPS's). MPS's are standards of MCCRES task performance; each standard is composed of various tasks and further divided into conditions and requirements. Requirements are specific actions which must be performed or behaviors which must be demonstrated in the accomplishment of a given task. Requirements may be amplified by key indicators (KI's) which are designed to provide specific, measurable actions or behaviors which must be present for the

requirement to be successfully completed and are graded 'yes,' 'no' or 'not accomplished.'

Potential areas for bias are explored, such as senior evaluator influence, other evaluator biases, and mission performance standards. In the study 243 requirements of selected MPS were examined. Fifteen (or 6.2%) were found to be susceptible to evaluator interpretation. Requirements containing phrases such as 'close attention' or 'processed with speed' cannot directly be answered 'yes' or 'no' and result in a determination based upon the evaluator's interpretation.

Field users of the MCCRES were interviewed to gain insight into potential problems. These Marine Corps officers of various ranks and military occupational specialties (MOS) were asked three questions relating to potential MCCRES evaluator bias. The results indicated bias was input through evaluator interpretation of performance criteria.

Three possible solutions to minimize bias are presented:

1. Evaluator Training. Errors will be reduced by training evaluators to be aware of the errors typically committed by evaluators and to ensure potential evaluators are well versed in the areas they are chosen to evaluate. The formation of a formal MCCRES evaluation team is proposed to ensure trained, knowledgeable evaluators. Other advantages such a team would provide, such as reduced training costs and more standardization of the evaluation base, are also discussed.

2. Evaluator Testing. Testing is recommended as a method of both controlling and controlling for evaluator bias. Testing could be used to select evaluators which demonstrated the least bias in their responses. Another use of testing would be development of a 'bias profile' which would be

controlling and controlling for evaluator bias. Testing could be used to select evaluators who demonstrated the least bias in their responses. Another use of testing would be development of a 'bias profile' which would be used to 'standardize' or normalize an evaluator's rating of an evaluated unit.

3. Quantification of MPS's. Reducing subjectivity in the mission performance statements would result in reducing evaluator bias.

TABLE OF CONTENTS

I.	INTRODUCTION	-----
II.	EVALUATION	-----
A.	DEFINITION AND PURPOSE OF EVALUATION	-----
1.	Definition of Evaluation	-----
2.	Purpose of Evaluation	-----
B.	PRINCIPLES OF EVALUATION	-----
C.	APPROACHES TO EVALUATION	-----
1.	The Systems Analysis Approach	-----
2.	The Behavioral-Objectives (Or Goal Based) Approach	-----
3.	The Decision-Making Approach	-----
4.	The Goal-Free Approach	-----
5.	The Art Criticism Approach	-----
6.	The Professional Review (Accreditation) Approach	-----
7.	The Quasi-Legal (Adversary) Approach	-----
8.	The Case Study (or Transaction) Approach	-----
9.	Summary	-----
D.	WHEN TO EVALUATE	-----
1.	Context Evaluation	-----
2.	Input Evaluation	-----
3.	Process Evaluation	-----
4.	Product Evaluation	-----
E.	SUMMARY	-----
III.	EVALUATORS	-----
A.	OBJECTIVITY	-----
B.	VALIDITY	-----

C.	ERRORS -----	
1.	Variable Errors -----	
2.	Constant Errors -----	
D.	EVALUATOR SOURCES -----	
1.	Superior Evaluators -----	
2.	Peer Evaluators -----	
3.	Disinterested Party Evaluators -----	
E.	ERROR SOURCES -----	
1.	Social Interaction -----	
2.	Evaluator Inexperience -----	
3.	Role Conflict -----	
4.	Evaluator Knowledge of Evaluation Purpose -----	
F.	ERROR REDUCTION TECHNIQUES -----	
1.	Evaluator Training -----	
2.	Dimensional Analysis -----	
3.	Testing Evaluators -----	
4.	Reducing Subjectivity of Evaluation Measures -----	
G.	SUMMARY -----	
IV.	MCCRES -----	
A.	APPROACH -----	
B.	STRUCTURE -----	
1.	Tactical Exercise Controller (TEC) -----	
2.	Evaluators -----	
C.	POTENTIAL PROBLEMS -----	
1.	Senor Evaluator Influence -----	
2.	Other Evaluator Biases -----	
3.	Mission Performance Standards -----	

D. POTENTIAL PROBLEMS PERCEIVED BY FIELD USERS	-----
E. RECOMMENDED SOLUTION	-----
1. Evaluator Training	-----
2. Evaluator Testing	-----
3. Quantification of MPS's	-----
F. SUMMARY	-----
REFERENCES	-----
DISTRIBUTION LIST	-----

LIST OF TABLES

- 3.1 Evaluator Ratings -----
- 4.1 MPS Requirements Susceptible to Evaluator Bias -----

LIST OF FIGURES

- 3.1 Evaluator Disagreements -----
- 4.1 MCCRES Evaluation Structure -----

I. INTRODUCTION

The Marine Corps Combat Readiness Evaluation System (MCCRES) was designed to provide timely and accurate information concerning the ability of active and reserve operating units of the Marine Corps to carry out assigned combat missions. The system uses "expert" evaluators from various specialty areas to observe and grade simulated combat operations. Aggregating these evaluations provides an overall view of a unit's readiness for combat. Feed-back from the evaluation allows the unit commander to identify and correct potentially problematic areas within his command.

Though the MCCRES is relied upon as a standard against which units are judged, is it possible that the readiness grade received could be more dependent upon the evaluator than the actual performance being graded? The purpose of this study is to examine the Marine Corps Combat Readiness Evaluation System in order to discover if the MCCRES is susceptible to biases which may cause the evaluations to inaccurately reflect the combat readiness of evaluated units. To guide the research, two specific questions were posed:

1. Can factors of the MCCRES evaluation which are subject to evaluator biases be identified?
2. How can these factors be controlled or controlled for?

These two questions were viewed from two major dimensions:

1. Evaluation--major approaches and principles.
2. Evaluators--sources and typical errors.

These dimensions were then related to the MCCRES and methods of controlling or controlling for evaluator bias were developed. A detailed literature search in the area of evaluation was conducted for this report. Also a sample of Marine Corps officers were interviewed for the study.

II. EVALUATION

This chapter addresses the evaluation process, presenting definitions, purposes and principles of evaluation, and explores some currently used approaches for conducting evaluations. The question of what to evaluate and when to evaluate are also investigated.

The term goal and objective are used throughout this and succeeding chapters. Goals refer to long range statements of purpose within the organization. They generally cannot be specifically stated and need not be attainable in the immediate future. Alternatively, objectives are more readily attainable in the short run and are specifically stated. They can appear as written statements which guide an organization's operations, and are a standard against which performance can be measured.

A. DEFINITION AND PURPOSE OF EVALUATION

1. Definition of Evaluation

There are many definitions of the term evaluation. Rather than select a single author's definition, two observations and two definitions of evaluation are presented here to show both the similarities and differences encountered in the field of evaluation research. These definitions and observations are given in order from simple to rigorous.

The first, more an observation than a definition, is from E. R. House:

At its simplest, evaluation leads to a settled opinion that something is the case. It does not necessarily lead to a decision to act in a certain way, though today it is often intended for that purpose....Evaluation leads to a judgment about the worth of something.

[Ref. 1: p. 18]

The second observation about evaluation, in particular the evaluation of a process, is that its scope "is confined to assessing what a particular program has accomplished in meeting its immediate objectives...", and assessing the "workability" of a program [Ref. 2: p. 11].

Reiken looks upon evaluation as "the measurement of desirable and undesirable consequences of an action that has been taken in order to forward some goal that we value" [Ref. 3: p. 54].

Finally, the definition presented by Stufflebeam et al., is that "...evaluation is the process of delineating, obtaining, and providing useful information for judging decision alternatives" [Ref. 4: p. 40].

There are two factors common to each of the preceeding observations and definitions. First, evaluation is concerned with making a judgment or assessment about something. Second, that judgment can be made in terms of some goal or objective. These two factors are used as a basis for a definition of evaluation developed in the next sections.

2. Purpose of Evaluation

Using the above descriptions of evaluation, the purpose of evaluation can be examined. Stufflebeam et al., stated simply that "The purpose of evaluation is not to prove but to improve" [Ref. 4]. Combining this statement with the ideas set forth above, we may look at evaluation as a judgment of something, say a program, with the purpose of improving the current attainment of that program's goals or objectives. Note that the judgment made may indicate some action which should be taken to improve the organization's goal attainment, but the judgment in and of itself does not cause the organization's goal attainment to improve. As such, the evaluation is a tool for program improvement. Evaluation as a tool for decision making is discussed by Anderson and Ball. Their use of the phrase "...to contribute to decisions..." [Ref. 5] in describing evaluation makes clearer the idea that evaluation is a tool rather than an end in itself.

If the above purposes of evaluation are accepted, then we may wish to form a new definition of evaluation. This definition takes into account

the purpose for the evaluation. Aggregating the previously cited authors' opinions and definitions we may look at evaluation as a judgment of some program with the purpose of contributing to decisions concerning the current attainment of that program's goals or objectives.

B. PRINCIPLES OF EVALUATION

There appears to be a general acknowledgement among authors of the evaluation literature that a group of principles exists which governs the conduct of evaluations. Tracey [Ref. 6] listed six principles which may be found in various forms in the writings of other authors [Refs. 1,4,5,8,9]. Evaluation must:

1. Be conducted in terms of purposes, that is, the objectives must be known. If the objectives are not known, the evaluation effort cannot measure how well they are being attained.
2. Be cooperative. Cooperation of all organizational levels is essential. Without free communication, evaluation results will not reach all parties, diluting the usefulness of the results.
3. Be continuous. Evaluation must be an ongoing process to accurately track performance and aid planning in light of current objective attainment.
4. Be specific. Generalizations are not as useful as specific information in providing performance information.
5. Provide means and focus to appraise self, practice and product.

The evaluation must provide information of sufficient quantity and specificity to evaluate not only the program output, but the mechanism of converting inputs to output and the individuals' performance within the mechanism.

6. Be based on uniform and objective methods and standards. Methods and standards which change from one evaluation to the next destroy trust and leave those being evaluated questioning how they should perform their work tasks.

[Ref. 6: pp. 14-15]

C. APPROACHES TO EVALUATION

How does one approach or categorize evaluation? The following section discusses eight approaches to or categories of evaluation forwarded by House [Ref. 1: pp. 21-43].

1. The Systems Analysis Approach

The systems analysis approach defines a small number of output measures and attempts to relate differences in programs to variations observed in the variables. The data acquired through this observation is quantitative in nature. Correlational analysis or other statistical methods are used to relate the output measures to the programs being evaluated. This method is widely used in the Department of Health, Education and Welfare in evaluating federal social welfare programs.

An example is the Office of Economic Opportunity (OEO) evaluation of the Neighborhood Health Center (NHC) program. The OEO defined five areas of interest to be investigated in determining the impact of the NHC's. These areas of interest were:

1. Success in the NHC's in providing comprehensive health care to the poor.
2. Patient reaction to the care received at the NHC's.
3. Degree of implementation of comprehensive and continuous family care at the NHC's.

4. Functional and organizational comparison of the NHC's.
5. Antipoverty consequences of NHC services.

[Ref. 7: pp. 107-121]

The NHC program was evaluated according to the attainment of the objectives which relate to the five specified interest areas.

One problem which may be seen with this approach is ensuring the output measures selected truly reflect the organization's goals. If the selected measures do not accurately reflect those goals, the outcome of this approach may be of limited use.

2. The Behavioral-Objectives (Or Goal-Based) Approach

This approach, popularized in business and government organizations as management by objectives, uses the stated objectives of a program as the output measure and evaluates program success by the attainment of these objectives. It can be seen that this method of evaluation addresses only the issue of program effectiveness, providing no information on program efficiency. In this sense, effectiveness is a measure of the extent to which an organization's objectives are achieved. Efficiency refers to the cost of converting program inputs to outputs, that is, the cost of objective achievement. An early advocate of this behavioral-objective approach was Tyler [Ref. 8] who advanced this method for evaluating educational goals in terms of student behaviors.

Peter F. Drucker popularized the term "management by objectives" in his book The Practice of Management [Ref. 9]. Implementation of Management by Objectives (MBO) forces individuals and organizations to define specific areas of responsibility in terms of measurable expected results, called objectives. Performance is determined by comparing objective attainment against the objectives stated. The popularity of the approach can be seen in its

widespread use. For instance, a 1976 study showed 41 percent of the hospitals surveyed used MBO and another 33 percent were planning to start in the near future [Ref. 10: pp. 8-11]. MBO is used not only as an evaluation approach, but as a means of planning, coordination, communication and control. An advantage is the explicit statement of objectives which lets workers know their specific duties and encourages communication between workers and supervisors relating to job performance. A major disadvantage is the problem of specifying behaviors rather than performance. Specific objectives are very measurable, but behaviors are not necessarily measurable in the context of contributing to goal attainment. Waks [Ref. 1: p. 487] argues that "...acting with purpose ..." is not equivalent to "...taking means to a well defined end." In other words, though a specified behavior may be observed, it does not follow that this behavior leads to a desired goal.

3. The Decision-Making Approach

As an earlier definition of evaluation implied, evaluation is closely related to decision-making. The decision-making approach holds that an evaluation is structured according to the decisions which must be made. It assumes that the decision-maker's concerns are the significant areas the evaluation must address. By structuring the evaluation in this manner, the results should be of greater use to the decision-maker. This approach relies heavily on survey methods such as interviews and questionnaires.

Stufflebeam et al. [Ref. 4], whose previously cited definition of evaluation includes the idea that evaluation is to provide information for judging decision alternatives, is an advocate of this approach in the field of education. The evaluation is structured with respect to the decision-maker's concerns and position in the organization, and specific evaluation subtasks are aggregated and communicated to the decision-maker in order to aid in the

decision process [Ref. 4]. This approach relieves the evaluator from having to guess the audience of the evaluation, thereby providing structure for the entire evaluation effort. On the other hand, this approach assumes that the decision-maker's goals are the same as those of the entire organization, which may or may not be the case.

4. The Goal-Free Approach

Each of the previously discussed approaches involved program evaluation in terms of program objectives and specific goals for the evaluation. The goal-free approach seeks to conduct evaluation in terms of program objectives without reference to the goals for the evaluation, indeed, the evaluator is purposely kept unaware of these goals so as not to be biased by them.

Scriven [Ref. 11], a leading proponent of this school of thought, feels that the goal-free approach is a valid method of reducing bias in evaluation, since knowledge of evaluation goals can influence the evaluator. For example, an evaluator who is tasked with conducting a performance evaluation of an employee with the explicit intent of determining whether the employee should be terminated may deliver a different evaluation if the intent is not stated. Evaluator knowledge that the evaluation may result in a worker being dismissed may bias the outcome of the evaluation. Being unaware of the evaluation intent may result in a more accurate representation of the worker's performance.

This approach is widely used in the area of consumer product evaluation. Various consumer organizations regularly evaluate products placed in the market without knowledge of the manufacturer's goals. These evaluations stress standards and criteria which the consumer organization feels are beneficial to the consumer. One main problem to overcome in this approach is the choice of evaluators. Scriven [Ref. 11] sees evaluators as experts,

able to eliminate and prevent both self-bias and bias of others from impacting on the evaluation. A variety of techniques, such as codes of ethics or double-blind experiments, are available to assist the evaluator in eliminating bias.

5. The Art Criticism Approach

This approach relies upon the critic to make judgment on a program much the same way an art critic would judge a fine painting. Though opinions on specific details may vary, there is generally a consensus among critics of a certain endeavor as to what constitutes a notable work. This implies an extensive base of common knowledge among those eligible to conduct such criticism.

Eisner makes a distinction between connoisseurship and criticism. While connoisseurship is "recognizing and appreciating the qualities of the particular" it requires no public disclosure or judgment. Criticism necessarily encompasses connoisseurship. "Criticism is the art of disclosing the qualities of events or objects that connoisseurship perceives" [Ref. 12: p. 197].

The key purpose of criticism is to increase awareness of a subject area and convey judgments in terms of criteria which are accepted among those knowledgeable in that area. It allows the uninitiated to gain an appreciation for that area through the critic's knowledge. Though generally associated with art, literature and other basically creative areas, the art criticism approach to evaluation has been applied to the field of education with some success.

A key problem with this approach is generating acceptance of the critic's criteria for judging a program. A critic may possess extensive knowledge in a particular field, but if the audience of the evaluation is not receptive, the criticism is not likely to carry much weight.

6. The Professional Review (Accreditation) Approach

The professional review approach has some distinct parallels with the art-criticism approach immediately above [Ref. 12]. Professional review relies upon expert opinion concerning generally accepted standards of performance in evaluating a particular area. The standards here, though, are usually more easily quantified, leading to a more structured approach in the evaluation. Professional review also is apt to use many members, organized as an accreditation or review board to conduct the evaluation. Standards and measurement criteria are determined by the professionals themselves since they are accepted as the experts in their fields. This approach produces an evaluation of professionals by professionals and its outcomes are not easily influenced by the layman.

7. The Quasi-Legal (Adversary) Approach

One of the long standing approaches for evaluating and policy-making is the quasi-legal approach. It is an approach to evaluation which closely imitates legal procedures. Information, or 'evidence,' concerning a program is obtained from 'witnesses,' much as testimony is received in a court of law. Information both for and against a particular program is presented, and great care is exercised to ensure that all pertinent information is received. A panel of evaluators then weighs the evidence heard and reaches a decision as to the worth of the program. Examples of this approach abound in today's government, ranging from local school board decisions for grade school curricula through presidentially appointed panels like the Warren Commission which investigated the assassination of President Kennedy .

This approach relies not only on expert evaluators as have several previous approaches, but it also encourages personal bias and opinion of those providing information. As Wolf notes:

The ultimate evidence which guides deliberation and judgment includes not only the 'facts,' but a wide variety of perceptions, opinions, biases, and speculations, all within a context of values and beliefs.

[Ref. 13: p. 21]

The ultimate goal of this approach is to reach a definite conclusion on some issue. Its conclusions address absolutes, such as 'Is the program meeting its goals' rather than matters of degree, as 'To what extent are our goals met.'

8. The Case Study (or Transaction) Approach

This approach is widely used and accepted in organizational studies. It focuses on program processes and interactions, both within and outside the program, with the intent of giving the reader of the case study a greater appreciation of the program's workings. This approach commonly presents interviews with people in the program and observations made by the interviewer at the program site in the form of a case. The case can be examined by evaluators and conclusions can be reached through discussions and sharing of ideas among the evaluators. The case study approach is used to increase understanding by illustrating how others view the program being evaluated. This approach helps the reader to understand the internal workings of the program and how program inputs are converted to outputs.

A major problem with this approach can be ensuring confidentiality for the members involved in the case study. Case study authors may have difficulty disguising all of the personalities involved in a case. Another problem which may be encountered is representing fairly the great diversity of actions and opinions which a large case study may entail. A complicated case

with many personal interactions can require an extensive editorial effort to ensure that it is accurate and understandable.

9. Summary

The above approaches are certainly not all inclusive. They are intended to show the variety of approaches available for conducting evaluations. Though the overall purpose of evaluation may be the same (i.e., providing information to aid in decision making) different situations may call for different approaches to provide the necessary information. The eight approaches demonstrated that techniques can be chosen which fit evaluation to the evaluator's skill (quasi-legal vs. professional review approaches), program objectives (system analysis vs. behavioral-objectives approaches), or even ignore evaluation goals (goal-free approach).

D. WHEN TO EVALUATE

Stufflebeam et al. [Ref. 4] provide a view of evaluation which investigates when in the program life cycle evaluation is to take place. They have defined four types of evaluation--context, input, process, and product--which serve functions from program inception through the final impact of the program on the system in which the program operates. Each evaluation type is explained briefly below.

1. Context Evaluation

Context evaluation is used in the planning process with the intent of identifying unmet goals or unused opportunities and identifying problems which prevent the goals from being met or the opportunities from being used. This problem identification leads to formulation of program objectives which are used as yardsticks against which program performance is measured. Stufflebeam

et al. [Ref. 4] further identify two modes of context evaluation: contingency and congruence. The contingency mode looks outside the system for factors which may yield improvements within. Typically, if-then type questions relating outside factors to objectives are asked--if our manning level is reduced by 20%, then can we carry out our mission? If research costs continue to rise, then is our present budget adequate? Congruence mode is a comparison between goals and actual information. This mode informs the organization as to its goal attainment. As opposed to contingency mode, congruence mode looks only within the system in question to provide evaluation data.

2. Input Evaluation

Input evaluation is concerned with the use of available resources in obtaining objectives formulated in context evaluation. Input evaluation is useful in providing information to be used in structuring the program. Besides program structuring, input evaluation also helps address such problems as the need for additional resources and other general strategic decisions.

3. Process Evaluation

Process evaluation begins after program approval and implementation. Process evaluation analyzes the program process as it is operating to provide information on whether the process is working as designed. Stufflebeam et al. [Ref. 4] point out that this type of evaluation is particularly important early in program implementation, when firm output information is not yet available. Process evaluation allows the organization to measure how well it is carrying out the program plan.

4. Product Evaluation

Product evaluation provides information on goal attainment, how well the stated objectives are met. Product evaluation is a major input to decisions which would modify the program after implementation.

The view provided by Stufflebeam et al. [Ref. 4] should not be regarded as an evaluation approach different from those listed by House [Ref. 1], but as an expansion of those approaches. Each of the eight approaches could be structured to look specifically at input, context, process or output. However, as implied earlier, the different approaches may not be equally effective in providing information in these four areas. The Stufflebeam et al. view can be seen as helping determine the timing of evaluations, indicating when to use one of House's approaches to provide information on specific portions of a program's life-cycle.

E. SUMMARY

This chapter has focused on the many ideas and approaches available in the evaluation literature. Definitions of evaluation and its purposes were presented to show the similarities and differences that exist in the evaluation literature. A definition of evaluation was formed. The definition looked upon evaluation as a judgment of some program with the purpose of contributing to decisions concerning the current attainment of that program's goals or objectives. Six principles for evaluation were also presented, demonstrating how and when evaluation should be conducted and what kind of information should be provided by the evaluation.

The basic concepts of evaluation were expanded by investigating eight approaches which are available to evaluators. These approaches provide different evaluation structures depending on the type of information desired or evaluation assets available. Finally, a view of evaluation which addresses when to perform evaluation was discussed.

With this grounding in the fundamental ideas of evaluation, the next chapter focuses on the evaluator's roles and responsibilities, and some problems associated with evaluation. The evaluator's implementation of the above principles and methods can greatly influence the eventual outcome of the evaluation.

III. EVALUATORS

The ideal rater who observes and evaluates what is important and reports his judgment without bias or appreciable error does not exist, or if he does, we don't know how to separate him from his less effective colleagues. [Ref. 14: p. 7]

Though the above statement may be true, many steps have been taken in evaluation to identify competent evaluators and improve performance of evaluators in general. This chapter looks at the evaluator, beginning with a discussion of objectivity and validity as they relate to evaluation. Who performs evaluations and whether they come from within or outside the organization is investigated. Advantages and disadvantages are presented for each evaluation source. A discussion of the kinds of errors evaluators typically make is presented along with sources which may cause these errors. The chapter closes with a discussion of several methods for reducing the amount of errors evaluators may bring into their evaluations, ranging from training the evaluator to improving the tools the evaluator uses in performing evaluation.

A. OBJECTIVITY

Objectivity, in the context of evaluation, is the ability to observe something only as it physically exists without the inclusion of personal feelings about the object. For example, the statement 'Joe is six feet tall' would be considered more objective than saying 'Joe is a Giant.' The former could be adequately demonstrated using a tape measure, while the latter is largely dependent upon the particular observer's concept of what is giant and

what is not. As House points out:

Objectivity is often equated with agreement among observers. Agreement is accomplished by having externalized specified procedures for observation. By this definition objectivity is achieved by having observers agree on what they see--replication of observation. [Ref. 1: p. 215]

House calls this the quantitative notion of objectivity. The concept of reliability in observation closely parallels this quantitative notion. Reliability is based on the ability to replicate observations. That is, if a particular observation is assumed to be reliable.

B. VALIDITY

The concept of validity is important to evaluation. If an observation does not accurately reflect the qualities of an object or construct (i.e., mental images) one wishes to measure, a 'true' evaluation of that object or construct may be impossible. Scriven [Ref. 15] describes the concept of validity as the qualitative sense of objectivity. He argues that, taken in the extreme, the quantitative notion of objectivity confuses the method of verification with 'truth.' An observation may be widely agreed upon and replicatable, but how closely does it represent the reality? How "good" is the observation? To illustrate, Scriven cited the incident of a television receiver evaluator observing picture quality. The evaluator used a mechanical device to measure decibel gain of the receivers, though there was little correlation between decibel gain and picture quality. The observations obtained were able to be replicated and the results widely agreed upon but they did not really relate to picture quality. In this case, the evaluation was quantitatively reliable but was not a "good" measure of picture quality

[Ref. 15]. This issue in evaluation--the goodness or quality of the measure--is commonly referred to as validity.

There are two general reasons that our measures are not totally valid: measurement deficiency and measurement contamination [Ref. 16]. Measurement deficiency occurs when the measure fails to take into account all of the factors present in our object or construct. For example, a measure of a data processing department's performance which accounted for quantity of output but neglected quality and timeliness would probably be considered deficient. Measurement contamination, in contrast to measurement deficiency, occurs when the measure takes into account factors which are not part of the object or construct. If our measure of the data processing department's performance includes items such as corporate sales or top management's perceptions of the department, the measure is likely to be contaminated. Both deficiency and contamination in the measurements of objects and constructs can adversely affect the usefulness of the measures.

C. ERRORS

There are a number of errors which evaluators may commit during the evaluation process. Cummings and Schwab [Ref. 16] discuss these errors in two main groups--variable error and constant error. These two groups are explained below, with examples.

1. Variable Error

Variable error is evaluator disagreement which manifests itself as differences in the scores of specific items of an evaluation. It may take two forms--disagreement between evaluators and disagreement over time.

a. Disagreement Between Evaluators

Suppose two evaluators, A and B, have observed five workers performing

their jobs and rated the workers' performance on a scale of 0 (poor performance) to 10 (high performance). The ratings are shown in Table 3.1. Note that there is total rating agreement only on worker 4 and the other ratings differ from 1 to 4 units.

TABLE 3.1
Evaluator Ratings

WORKERS	<u>RATINGS</u>	
	EVALUATOR A	EVALUATOR B
1	5	3
2	7	8
3	3	7
4	9	9
5	4	0

Taking the ratings obtained from A and B, we now wish to plot the scores, with evaluator A's rating representing the X-component of our plot and evaluator B's ratings representing the Y-component of the plot. The result is a graph as shown in Figure 3.1. The straight line extending from the origin and rising from left to right represents total agreement between the evaluators. The distance of each worker's score from the total agreement line is a measure of the disagreement between the evaluators. A linear correlation coefficient may be calculated which expresses the amount of agreement between the evaluators. Values for the linear correlation coefficient may vary from -1.0 (highly negative correlation, meaning that high values for the X-component tend to go with low values for the Y-component and low values for the X-component tend to go with high values for the Y-component) to +1.0 (highly positive correlation, meaning that high values for the X-component tend to go with high values for the Y-component and low values for the X-component tend to go with low values for the Y-component), with a value of

0.0 indicating no correlation (no predictable pattern). In this example, the linear correlation coefficient is 0.6 indicating some positive correlation between evaluators A and B. In general, a value in the range of 0.8 to 0.9 would tend to indicate a strong correlation between A and B. High correlation demonstrates reliability but does not guarantee a valid rating. It simply shows that A and B agree on what they have observed. Both A and B may be wrong in their ratings of worker 4, but their agreement would provide some confidence that their rating was correct.

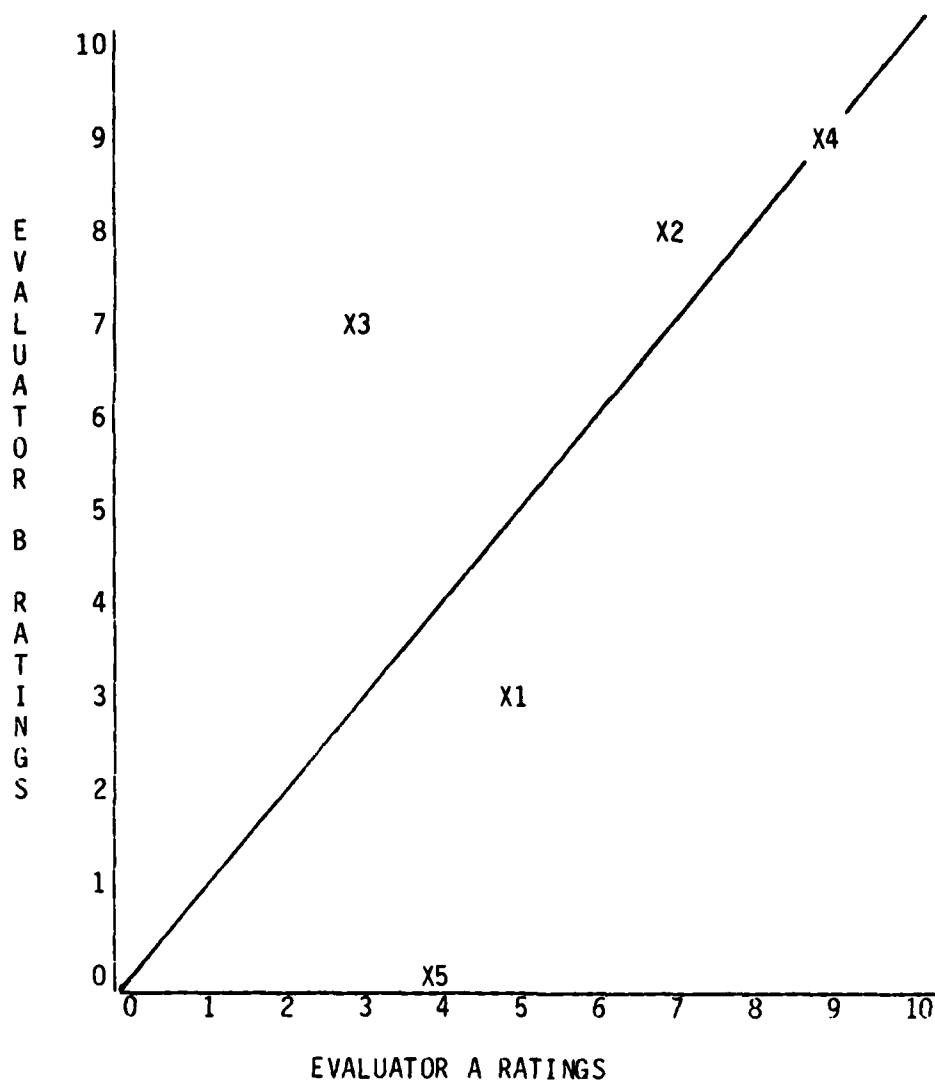


Figure 3.1 Evaluator Disagreements

Two methods which can reduce disagreement between evaluators are reduction or elimination of subjectivity in measurement instruments and ensuring evaluator familiarity with the job being evaluated. The former method reduces disagreement by relieving the evaluator of interpreting subjective measures. By using more objective evaluation measures, evaluator bias is less likely to be accidentally introduced [Ref. 20: p. 46]. Ensuring evaluator familiarity with the job being evaluated increases the likelihood of evaluating job factors which correlate highly with job performance.

b. Disagreements Over Time

Disagreements over time pertain to disagreements in evaluations made by one evaluator at different points in time. Suppose that, in the example of disagreements between evaluators, evaluator A's ratings represented an evaluation performed by A at time 1 and that evaluator B's ratings represented an evaluation performed by A at time 2. Calculation of the linear correlation coefficient would then measure how well evaluator A's ratings agree over time. However agreement of the ratings over time may or may not be appropriate. The reason for this is that differences in evaluations made at different points in time may be due to performance improvement or degradation of those being evaluated. A method for reducing disagreements over time, discussed later, is testing potential evaluators and choosing those who demonstrate little of this error.

2. Constant Error

Where variable errors tend to create differences between evaluations, constant errors tend to cause spurious similarities. Constant error takes three forms--halo error, central tendency and leniency.

a. Halo Error

Halo error occurs when an evaluator fails to differentiate among individual items or dimensions in the evaluation, but evaluates on the basis of overall impression. The boss who observes only an employee's written work but rates the employee high in areas such as initiative and personal relations has made a halo error.

b. Central Tendency

Central tendency is the tendency for evaluators to rate all dimensions of an object near the middle of the evaluation scale, avoiding the extremes.

c. Leniency

This error is committed when an evaluator tends to rate all objects too high. The "easy grader" consistently delivers inflated rating marks. The opposite error, that of rating all objects too low is called strictness.

Evaluator training in the area of constant error is a useful technique in reducing these errors. A discussion of this technique is presented in a later section.

D. EVALUATOR SOURCES

Evaluators may come from many places within and outside an organization. Though evaluations by superiors are very common, alternative sources of evaluation exist--peer, subordinate, self and disinterested party or outside evaluators.

1. Superior Evaluators

Evaluations by superiors are a widely used method in today's organizations. Superiors are chosen for many reasons: job experience, familiarity with subordinate positions and job skills--even tradition. Superiors are often the logical choice as evaluators, their position in the organizational hierarchy is such that they determine to a great extent the incentive and

reward system for their subordinates. As such, their evaluations of subordinates may lead to direct reward or punishment without passing through another level of hierarchy and this immediate evaluation-incentive tie keeps subordinates appraised of their performance.

Some problems can exist with supervisor evaluations. First, if the subordinate being rated does not work directly for the evaluating superior or if there is substantial physical separation of the supervisor from the subordinate, the supervisor's observation of the subordinate's job performance may be limited. Also, due to rapidly changing technology, the superior may not have enough understanding of the subordinate's actual on-the-job responsibilities to adequately rate the individual's performance.

2. Peer Evaluators

Peer evaluators are those individuals who work at the same organizational level as the person rated. Many organizations avoid using peer evaluations, dismissing the technique as a "popularity contest." Peer evaluator-evaluated friendship is seen as biasing the validity of this technique. This may be due to the perception that friends tend to minimize or overlook one another's shortcomings and only elevate good points, or mistake pleasing personal attributes for indicators of high job performance. However Klimoski and London [Ref. 17] and Love [Ref. 18] have shown that evaluation validity is not significantly affected by friendship bias, and that in some circumstance, peer evaluation appears to offer great benefits to an evaluation program.

3. Disinterested Party Evaluators

Disinterested parties can possibly be found within the organization or outside. They may come from any organizational level so long as they have no vested interest in the outcome of their evaluations. Some organizations bring

in outsiders to perform this function, feeling that lack of personal contacts within the organization facilitates a more objective evaluation.

A problem which may occur with disinterested party evaluators is that, aside from having no vested interest in the evaluation outcome, they may have limited insight into the factors which indicate good job performance. Outsiders brought in to perform evaluations may not fully grasp factors such as organizational politics and interpersonal relationships which can greatly influence overall job performance.

As Holzburg [Ref. 19] has shown, the source of the evaluator affects what factors or dimensions are chosen for evaluation. For instance, if one examined the broad area of secretarial job performance, many individual dimensions could be identified, such as typing speed, typing accuracy, shorthand ability, organization, ability to speak effectively on the telephone and many others. Unless the list is complete, it will be deficient in measuring secretarial job performance. Holzburg found that evaluators from different sources chose different sets of dimensions for evaluation. Continuing the secretarial job performance example, consider an evaluation of performance performed by a worker's superior and a peer. The superior may rate the worker's clerical performance according to how many pages are typed per hour--assuming, perhaps incorrectly, that quantity of pages typed also indicates quality. The peer, who must correct any errors made by the worker, may be concerned with quality of output. What this means to the individual being evaluated is that performance grades received may be due more to the source of the evaluator than the job performance.

E. ERROR SOURCES

Many factors contribute to evaluator error. Though often grouped under the general heading of bias, specific factors have been investigated as a way of ensuring objective and valid evaluations. This section looks at several of the factors contributing to evaluator error, and the next section discusses some methods suggested for reducing these errors.

1. Social Interaction

Social interaction, or friendship bias, is cited as a reason for avoiding peer or superior evaluations. As previously noted, this bias is thought to adversely affect peer evaluations. This bias is also seen in superior evaluations. However, superiors and peers are used as evaluators. Their use as evaluators--at least for the organizations that use them--would indicate that social interaction as a source of error is not considered so severe as to disqualify peers or superiors as evaluators.

2. Evaluator Inexperience

Evaluator inexperience and lack of training in evaluation procedures tend to contribute to halo and leniency errors [Ref. 20]. Poorly defined measures force the inexperienced evaluator to make interpretations which, due to limited background, may not accurately reflect performance. Closely associated with this idea is the evaluator's effectiveness on the job. Low evaluator effectiveness correlates strongly with low evaluation accuracy.

3. Role Conflict

A factor contributing to evaluator error is the role conflict experienced by many evaluators. Dayal has noted:

The manager has to accept the responsibility to judge the performance of other people. Often this responsibility is hesitantly taken because he feels uncomfortable in his role as a judge. [Ref. 21: p. 29]

One effect of this evaluator discomfort is that evaluation results tend to group near the upper end of the rating scale [Ref. 21]. A possible reason for this effect is that giving low ratings may result in slower promotion or even firing of an employee, for which the evaluator giving the ratings may feel responsible. Ratings at the high end of the scale reduce the probability that employees will experience layoffs or slower promotion and the evaluator will feel less responsible if such actions do occur.

4. Evaluator Knowledge of Evaluation Purpose

As previously stated, Scriven [Ref. 11] has suggested that evaluator knowledge of the evaluation purpose may be another nonperformance factor influencing the actual performance rating received. A study by Gallagher [Ref. 22] investigated whether ratings of performance varied when evaluators were given different purposes for the evaluations. The results support Scriven's contention. Gallagher's discussion of the results concludes "...that a single performance evaluation should not be used for different purposes since the stated purpose of the evaluation can affect the actual performance rating" [Ref. 22: p. 38].

F. ERROR REDUCTION TECHNIQUES

Many techniques are available to help reduce evaluator error. These techniques have been investigated by various evaluation researchers (e.g., Bernardin [Ref. 23], Wiley and Jenkins [Ref. 24], and Scott [Ref. 20]) and some suggested solutions are presented here.

1. Evaluator Training

Bernardin in a study of comprehensive vs. abbreviated evaluator training programs found that evaluators with comprehensive training performed better than evaluators with abbreviated training in terms of leniency error and halo effect [Ref. 23]. In the study comprehensive training was a one hour session

consisting of definitions, graphic illustrations and examples of halo, leniency and central tendency errors which were presented to students who were acting as evaluators of peer performance. The trainees were also given data to evaluate for errors and the evaluations were discussed. Abbreviated training was a five minute session with definitions of the error types and a single illustration of each.

The results of this study indicated that the quality of evaluation by those who underwent comprehensive training was superior to those who received abbreviated training at the first rating period, and both training groups were superior to the control (untrained group). Another result was that the positive effects of the training program were virtually nonexistent after one additional rating period [Ref. 23]. One might argue that for an organization contemplating a training program for supervisory personnel the above information may indicate that a comprehensive training program would lead to fewer evaluator errors than an abbreviated training program. As the effects of both training programs tend to rapidly diminish with time, however, a shorter training program regularly administered may deliver more positive effects in the long run.

2. Dimensional Analysis

As discussed previously, different evaluators from different sources perceive performance in different ways. To account for this, the various dimensions of subjective evaluation areas should be identified. Once the various dimensions are identified, the different combinations of dimensions used by various evaluators should then be analyzed. Such an analysis can provide insight into the particular concerns of evaluators from various sources. Klimoski and London [Ref. 17] argue that supervisors may be less

able to discriminate between items related to competence and those related to effort, whereas nurses rating themselves and peers can make that distinction. This would suggest that supervisors are more likely to consider effort as an indicator of competence than peers. By accounting for the dimensions used by various evaluators from various sources, dimensional analysis can allow performance measures to be tailored according to the anticipated evaluator source, or it may be used after the fact to help explain ratings received in particular areas.

3. Testing Evaluators

Wiley and Jenkins [Ref. 24] had 109 Air Force navigator students estimate qualifications needed to perform various Air Force tasks using an experimentally standardized task list and sets of five rating scales. Their estimates were aggregated and a consensus or pooled estimate group was formed. These students, after one month, again estimated qualifications and the students were scored by correlating their estimates with the key of pooled estimates. The study showed that evaluators who tended to agree with the consensus also tended toward agreement with the consensus in later evaluations [Ref. 24].

The above findings tend to suggest that a standardized test could be developed to rate potential evaluators. A consensus key which corresponds to the organization's view of performance would make it possible to select evaluators with corresponding views. This would help ensure organizational goals are being pursued by the evaluation process.

4. Reducing Subjectivity of Evaluation Measures

Performance appraisal systems are commonly regarded as being too subjective in nature, relying primarily on human judgment for gathering information pertaining to measures [Ref. 20]. Elimination of all factors

which cannot be objectively measured would naturally lead to minimal subjectivity. While this elimination may or may not be possible, it is possible to develop a system where the evaluator reacts to stimuli which are relatively free of subjective or irrelevant influences rather than stimuli which require the evaluator's judgment [Ref. 16: pp. 89-92]. The stimuli take the form of actual on-the-job incidents which the evaluator simply observes without interpretation. These incidents, or 'critical behaviors,' represent actions normally associated with clearly successful or unsuccessful task performance. The evaluator in this role acts as a reporter of actions rather than a judge who values actions [Ref. 20].

One problem associated with this method is the choice of critical incidents or behaviors. Some person or group of people must be designated to decide what incidents are to be used in evaluation. The selection of the individuals may cause bias or errors in the identification of the incidents.

G. SUMMARY

This chapter has investigated the evaluator as part of the scheme of evaluation. The concept of objectivity and validity were introduced and explained as they pertain to evaluation. Sources of evaluator error were then discussed. Evaluator errors were divided into variable and constant errors, and each of these areas was broken into specific error types. Evaluator sources--superior, peer and disinterested party--were discussed with advantages and disadvantages of each source considered. A discussion of error sources, along with techniques to reduce these errors closes the chapter. The last section suggests that training and testing evaluators and taking measures to reduce the subjectivity of evaluation measures can reduce an evaluator error.

IV. MCCRES

The purpose of the Marine Corps Combat Readiness Evaluation System (MCCRES) is to provide a timely and accurate evaluation of the readiness of Fleet Marine Forces, including Reserve units, to accomplish assigned missions. [Ref. 26: p. I-A-1]

To achieve the objective of timely and accurate readiness evaluation, the MCCRES has been designed to allow observation of Marine units in simulated combat situations. The MCCRES promotes use of a standardized evaluation process and reporting system to provide feedback to the evaluated unit indicating strengths and weaknesses in a combat readiness posture. Building upon Chapters II and III, this chapter focuses on the evaluation process in an attempt to identify areas where evaluators may commit errors or inject bias into the evaluation. The general evaluation approach and structure of the MCCRES are discussed first, followed by an investigation of potential sources of error. The final section discusses some solutions to minimize the effects of evaluator bias.

A. APPROACH

The MCCRES approach to evaluation may be compared with the Professional Review (Accreditation) Approach discussed by House [Ref. 1]. It is an evaluation system conceived within the Marine Corps, graded by Marines and using standards developed by Marines. As such, it closely parallels the Professional Review Approach. In this approach, a particular profession sets standards of performance for itself and conducts internal evaluations. The reasoning for the internal evaluations is that members of that profession are considered experts in that field.

In choosing evaluators to perform MCCRES evaluations, it is desirable that evaluators have recently served successfully in a billet relating to the

function they are to observe. This means, for example, that a Rifle Company evaluator should have recently served successfully as a Rifle Company commander. Successful recent billet performance increases the probability that evaluators will recognize adequate mission performance.

B. STRUCTURE

The MCCRES evaluation structure can be depicted as a four-tiered hierarchy as shown in Figure 4.1. Of particular importance to this discussion are the bottom two layers--the Tactical Exercise Controller (TEC) and the Evaluators. It is here that mission performance is observed, analyzed and reported.

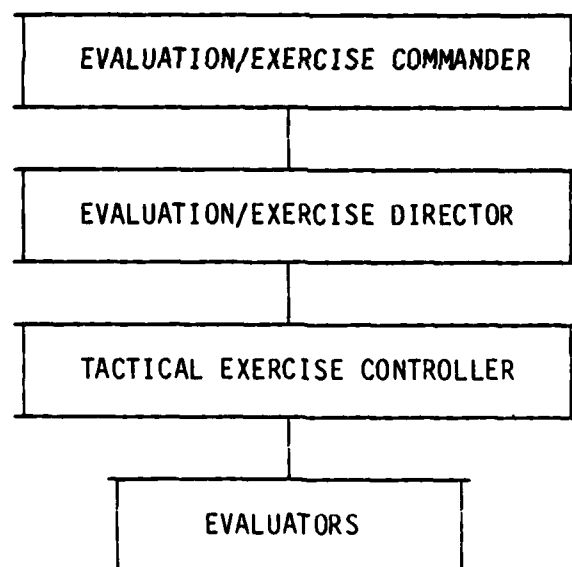


Figure 4.1 MCCRES Evaluation Structure

1. Tactical Exercise Controller (TEC)

The TEC compiles and analyzes the results of the evaluations which have been submitted via the evaluator's data sheets and submits a formal report to the Exercise Director. Among the TEC's duties and responsibilities are determination of specific Mission Performance Standards to be tested, extensive and detailed training of evaluators, development and control of intelligence play throughout the problem, and organization of the Tactical Exercise Control Group to plan and conduct the exercise. The TEC relies on the evaluators to report exercise progress and mission performance of the evaluated units. The former information is received primarily via radio communication while the latter arrives in the form of evaluator data sheets.

2. Evaluators

Evaluators have three main roles in the MCCRES:

1. Exercise controllers to ensure the exercise proceeds as planned.
2. Umpires to resolve disagreements between exercise and aggressor forces.
3. Performance evaluators to observe task performance as related to Mission Performance Standards being graded.

As an exercise controller, evaluators work as an extension of the will of the TEC. They may increase or decrease the operational tempo of the problem through the use of such items as aggressor forces, intelligence reports or simulated fires. They may create situations which require reaction by the evaluated unit by insertion of prescribed events into the play of the tactical problem. Action observed at this level is provided to the TEC primarily by radio to assist the TEC in determining if the exercise pace is satisfactory.

As umpires, evaluators are tasked with resolution of disagreements which may occur between evaluated units and aggressor forces. For example, if an

evaluated unit was ambushed by an aggressor force, an evaluator would make a determination as to the outcome of the ambush and assess casualties accordingly.

In the role of performance evaluators, evaluators observe unit performance of prescribed tasks and make a determination as to the unit's ability to satisfactorily carry out the task. These determinations are recorded as "YES," "NO" or "NOT APPLICABLE" marks on the evaluators data sheet. A mark of "YES" denotes that all facets of a particular requirement were met. Conversely, a "NO" mark shows that all portions of a requirement were not met. "NOT APPLICABLE" areas are those not tested or which do not apply to the scenario at hand.

Each unit evaluated has a senior evaluator who conducts a post exercise wrap-up and compiles the data sheets from all subordinate evaluators. At this wrap-up, resolution of each "YES," "NO" and "NOT APPLICABLE" rating is made for each requirement tested. The resolution of the evaluators' data sheets result in "YES," "NO" or "NOT APPLICABLE" ratings for each requirement as it pertains to the entire unit. The senior evaluator provides the data sheets to the TEC for compilation and further use by the TEC. An assessment of "COMBAT READY" or "NOT COMBAT READY" for the entire unit is also passed to the TEC by the senior evaluator.

The senior evaluator's relationship with other evaluators is a senior-subordinate type. Senior by position and generally by military rank, the senior evaluator is in charge of the evaluation team and is responsible for evaluating the performance of the entire unit being evaluated. The senior evaluator is appointed by name by the Exercise Director (an officer senior to the commander of the organization being evaluated) and as such, maintains an independent relationship to the organization being evaluated. Other members

of the evaluation team, subordinate to the senior evaluator, are responsible for evaluating the subordinate units (both organic and attached) and other organizational functions (such as command and control and fire support coordination) of the overall unit being evaluated.

3. Mission Performance Standards

Mission Performance Standards (MPS's) are standards of task performance used in MCCRES. Each standard is composed of various tasks. For example, the MPS Continuing Actions By Marines is composed of twelve tasks such as Discipline, Dispersion, Security and Casualty Handling. These tasks are further divided into conditions and requirements. Conditions specify the circumstances under which requirements must be performed and provide recommendations to the evaluator concerning time and space limitations which may be imposed on the evaluated unit. Requirements are specific actions which must be performed or behaviors which must be demonstrated in the accomplishment of a given task. Requirements which may need further information to guide evaluators in the determination of satisfactory performance are provided with Key Indicators (KI's) of performance. KI's are an attempt to provide an objective foundation upon which an evaluator can base judgment of satisfactory requirement performance. They are designed to provide specific, measurable actions or behaviors which must be present for the requirement to be successfully completed.

Consider the KI for the requirement Weapons Maintenance Discipline. "Marines take care to clean their weapons, both individual and crew served, daily. Weapons are safeguarded. Care of weapons enforced by leader." The KI tells what is to be done (clean weapons, both individual and crew served), when it is to be done (daily), who does it (Marines), and who supervises

(leaders). KI's for other requirements provide similar types of information to make requirements more objectively measurable by the evaluator.

C. POTENTIAL PROBLEMS

This section discusses the areas in which evaluators may inject bias into the MCCRES. The discussion is presented in three parts: Senior evaluator influence, other evaluator bias and MPS problems. Some general solutions to these problems are suggested here with more specific solutions presented in the following section.

1. Senior Evaluator Influence

The senior evaluator can inject bias in two major ways. First, as the senior member of the evaluation team, the senior evaluator sets the tone for the other evaluators. If the senior evaluator projects a hard-line, "by the book" approach toward the evaluation, evaluators may tend to view task requirements with little flexibility. On the other hand, in a situation where the senior evaluator projects a less rigorous attitude toward the evaluation, evaluators may tend to view task requirements less rigidly. As a result of evaluator perceptions of the senior evaluator's wishes, the evaluation delivered may be biased.

The second major way in which the senior evaluator may inject bias is in the resolution of other evaluators' ratings. These ratings are obtained from the data sheets of the other evaluators. The senior evaluator depends upon the observations made by the other evaluators to provide data which accurately reflects the performance of the entire unit. Depending on the senior evaluator's perceptions of the other evaluators' competence and the senior evaluator's own perception of successful task completion, the senior evaluator's data for the TEC may or may not accurately reflect the overall

unit's abilities. As an example, suppose an infantry battalion conducted an attack on an aggressor force and that two of the companies performed extremely well while one company performed poorly. If, in the senior evaluator's opinion, the offending company's performance was not critical to the entire unit's mission performance, a rating of "YES" could be delivered for the battalion for the task "ATTACK" as it pertains to the entire unit [Ref. 26: p. I-C-8]. On the other hand, if the senior evaluator thought the one company's performance was such that it negated the accomplishments of the other two companies, a rating of "NO" could conceivably be returned for the battalion for the task "ATTACK" as it pertains to the entire unit. The senior evaluator made a decision based on personal judgment, possibly reflecting the unit's mission performance inaccurately.

2. Other Evaluator Biases

The evaluators who observe task performance and report to the senior evaluator are presented with a continuing opportunity to inject bias into the MCCRES. The discussion of the areas where these evaluators may inject bias is organized in two groups: errors and evaluator sources.

a. Errors

Evaluator bias manifests itself as any deviation from the objective 'truth' concerning an evaluated unit's performance. In this respect, bias may be regarded as an error of leniency, strictness or halo effect. The first two errors result in ratings which are respectively too "easy" or too "hard," while the last error tends to cause ratings to group around one value on the rating scale. To illustrate, consider an evaluator rating the requirement Equipment Maintenance. The first portion of the KI for this requirement states "Vehicles, generators, etc., are given close attention by the Marine assigned to operate them" [Ref. 26: p. II-A-6]. The lenient evaluator may

consider visual observation every four hours constitutes close attention, while a strict evaluator considers maintenance conducted every other hour as an indicator of close attention. If a Marine is observed by these two evaluators checking the assigned equipment at strict four hour intervals because that is what the operating manual calls for, the Marine will receive a different rating from each of the evaluators. In this case, the second evaluator has injected bias by committing the error of strictness.

As an illustration of halo error, suppose an evaluator is rating a unit on a task which contains five requirements. At the outset of the observation period, the unit was particularly outstanding in carrying out the first requirement. Based upon the outstanding performance, the evaluator expects similar performance for the other requirements of the task. Such expectations may influence the evaluator to "see" only outstanding performance. Mistakes and poor performance are viewed with the attitude that "...they really know better, they just weren't paying attention today..." As a result of this attitude, a "YES" rating is delivered for the entire task, even though not all requirements were successfully completed. This evaluator has committed a halo error since the rating has been influenced by the outstanding performance of only one requirement of the entire task. It must be noted that this error can also be observed in the opposite sense, that is, a particularly bad observation can bias the evaluator to view an entire task unfavorably.

b. Evaluator Sources

In the discussion of the three main sources of evaluators--superior, peer and disinterested party--it was shown that they vary greatly in perceptions of task performance. This difference in perception is related to the dimensions of the task being evaluated. In the context of MCCRES this means that superiors may not perceive task performance in the same way as peers. The

last evaluation source, the disinterested party, brings with it the additional potential problem of not understanding the process being graded.

Many of the potential problems associated with evaluators from different sources are diminished by two criteria for MCCRES evaluators. The first criterion is that evaluators should have recently served a successful tour in a billet related to the one they are evaluating. A key word in this stipulation is recently. An evaluator who has recently served in a billet similar to the one being evaluated is more likely to recognize those task dimensions which indicate successful task performance than an evaluator who has not recently held such a position.

In addition to the problem associated with evaluators from the alternative sources identifying varying dimensions, social interaction between sources and the evaluated unit can be problematic. Both seniors and peers within an organization tend to interact in formal as well as informal ways. This informal or social interaction may be carried into the evaluation as a bias. The second criterion is that "...it is desirable that evaluators be obtained from adjacent commands not directly related to the organization being evaluated" [Ref. 26: p. I-C-9]. This may result in a reduction of bias created by social interaction. This reduction is due to decreased daily interaction between members of adjacent units as compared to daily interactions among members of a single unit.

3. Mission Performance Standards

All of the evaluators from all the sources have at least one thing in common: they use the Mission Performance Standards to evaluate unit combat readiness. A potential problem associated with the MPS's is their subjectivity. This subjectivity permits evaluator interpretation of standards which may result in biased evaluations.

In an attempt to determine the extent of the MPS's subjectivity, the re-quirements for the MPS's of Continuing Actions by Marines, Command and Control and Fire Support Coordination were examined. The criterion used to determine the subjectivity of a requirement was the ability of the requirement to be quantified. If the requirement was expressed in terms which are physically measurable, such as units of time or distance, then it was considered objective. Requirements containing phrases which require interpretation by the evaluator, such as "close attention," were considered subjective. The meaning of the subjective requirements can depend upon the evaluator's interpretation of the requirement's wording.

Of the 243 requirements for the above MPS's, 15 were found to be susceptible to evaluator interpretation. This is approximately 6.2 percent of the requirements for these three MPS's. These 15 requirements contain phrases such as "close attention" or "processed with speed" to describe satisfactory requirement performance. Without clear guidance as to what constitutes close attention or processed with speed, different evaluators may interpret the requirement to have different meanings. This difference in interpretation means that two evaluators observing a particular requirement being performed could return different ratings of requirement performance, depending on how the requirement is interpreted. For each of the 15 requirements, the requirement number and the subjective phrase contained in the requirement is listed in Table 4.1

TABLE 4.1

MPS Requirements Susceptible to Evaluator Bias

<u>Requirement Number</u>	<u>Subjective Phrase</u>
2A.1.1.3	"close attention"
2A.1.1.4	"orderly and organized fashion"
2A.1.1.7	"exhibit restraint"
2A.1.1.8	"light use to a minimum"
2A.1.8.6	"COMSEC material safeguarded"
2A.1.11.14	"processed with speed"
2A.2.7.2	"provided with security"
2A.2.8.2	"safeguards classified material"
2A.2.9.5	"neat and orderly"
2A.2.9.6	"dispersed to reduce vulnerability"
2A.2.10.5	"dispersed"
2A.3.4.5	"closely monitors"
2A.3.4.7	"timely manner"
2A.3.5.3	"accurate plots"
2A.3.5.7	"closely monitors"

D. POTENTIAL PROBLEMS PERCEIVED BY FIELD USERS

In an attempt to gain some insight into potential MCCRES problems as perceived by users in the field, a small sample of Marine officers was interviewed. Six officers who were students at the Naval Postgraduate School and ranged in grade from O-2 to O-4, representing MOS 0302 (Infantry Officer), MOS 1302 (Engineer Officer), MOS 7562 (Pilot HMM CH-46) and MOS 7587 (Airborne Radar Intercept Officer, F4N/J/S) were interviewed. The interview consisted of three questions:

1. Do you feel that an evaluator can affect a MCCRES evaluation through personal bias?
2. How is this bias input?
3. In what area do you feel bias is most likely to occur?

The results of these interviews demonstrated that there was close agreement on each of the questions across both MOS and grade. All interviewees felt that an evaluator could affect a MCCRES evaluation through personal bias. This

bias was seen as being input through evaluator interpretation of performance criteria. These criteria take the form of task requirements. Responses to the last question indicated field users felt bias is most likely to occur in those areas to which numerical measures are not easily attached. They said that areas which lend themselves to quantifiable measurement are less likely to contain evaluator bias than non-quantifiable areas. The potential problems with MCCRES, as perceived by the sample of field users, are a subset of the potential problems discovered through analysis of the MCCRES. Though the sample of six officers is small, the unanimity of their views indicates that evaluator bias may exist in MCCRES evaluations.

E. RECOMMENDATIONS

The problems discussed in the previous two sections demonstrate the variety of ways in which an evaluator may introduce bias into a MCCRES. In order to minimize biased input, three possible solutions are presented: evaluator training, evaluator testing and quantification of subjective MPS requirements.

1. Evaluator Training

As previously noted, evaluator training has proved to be an effective tool in reduction of evaluator error. Bernardin [Ref. 23] demonstrated that evaluators receiving comprehensive training showed greater error reduction results than evaluators receiving limited training. Both of these groups showed less error than evaluators who have received no training.

Current MCCRES standards task the TEC with conducting extensive and detailed training of evaluators. In the experience of the officers attending the Naval Postgraduate School, who were questioned concerning evaluator training, the training is geared toward introducing the evaluator to the exercise scenario with no specific mention of the errors which evaluators

typically commit. By making MCCRES evaluators aware of the errors typically committed by evaluators, the MCCRES evaluators are less likely to commit these errors, reducing biased input. An evaluator training package addressing both scenario development and possible evaluator error should be created to more fully exploit the potential of comprehensive evaluator training outlined by Bernardin [Ref. 23].

Another aspect of evaluator training is ensuring potential evaluators are well-versed in the areas they are chosen to evaluate. Choosing knowledgeable evaluators tends to increase the probability that those factors which indicate successful task performance are considered during the evaluation.

One method to ensure trained, knowledgeable evaluators for MCCRES evaluations would be the formation of a formal MCCRES evaluation team. By choosing team members who have demonstrated proficiency in their MOS's and evaluation techniques through training, a skilled cadre of evaluators could be assembled. Some of the advantages of forming a formal MCCRES evaluation team would be minimization of evaluator training costs, minimization of social interaction with evaluated units and a more standardized evaluation base. Evaluator training costs would be minimized since the same evaluators would be frequently used. Though training effects diminish rapidly with time, re-training for each successive evaluation could demonstrate a learning curve, reducing costs over time. Social interaction would be minimized due to lower daily contact with evaluators, as opposed to the interaction which occurs among adjacent commands. The last factor, standardization of the evaluation base, results from the continuity of the formal evaluation team.

A MCCRES evaluation team could be composed of personnel from units such as Division Schools, or it could reside outside the active duty forces at a Reserve unit, since the MCCRES is to evaluate both active and reserve forces.

Having reserves evaluate MCCRES would also offer the additional benefit of keeping the reserve up to date and strengthening the tie between active and reserve forces in the Marine Corps.

2. Evaluator Testing

Evaluator testing is a method of both controlling and controlling for evaluator bias. In the former case, a test can be constructed which would indicate the areas in which a prospective evaluator demonstrates bias. By testing a number of these prospective evaluators, those who demonstrate little or no bias could be chosen to conduct MCCRES evaluations, thereby minimizing the likelihood of evaluator bias input. For instance, consider a test in which evaluators are graded according to their agreement with an answer key. Further, suppose the answer key is composed of the pooled answers of a group of unbiased evaluators. As suggested by Wiley and Jenkins [Ref. 24: p. 217], evaluator agreement with the key can be used to predict the likelihood of evaluator bias. Those evaluators showing close agreement with the key of "unbiased" answers could then be chosen to perform evaluations.

The same test, analyzed differently, can be used to control for evaluator bias. For instance, the results of the test are analyzed to discover in which areas an evaluator's biases exist. From this analysis a bias profile could be constructed which could allow evaluation results to be "standardized." For example, assume a MCCRES evaluator's bias profile showed significant deviation toward strictness in the area of discipline. Assume that during the conduct of a MCCRES evaluation the senior evaluator notes this evaluator's data sheet has a "NO" rating for many of the requirements of the task DISCIPLINE. The senior evaluator, knowing that this evaluator tends to be particularly strict in evaluating discipline, may wish to obtain additional performance

information concerning the unit evaluated, since the evaluator's ratings may not accurately reflect the unit's actual performance.

3. Quantification of MPS's

The last suggested method of controlling evaluator bias is quantification of subjective MPS requirements. Quantification of evaluation measures, as Scott [Ref. 20] suggests, reduces the evaluator's task from interpreting the evaluation measure, in this case the MPS requirements, and comparing task performance with this interpretation to reporting whether task performance meets the requirements. For example, instead of trying to decide how fast "process with speed" is, reporting whether the unit was able to "process within two hours" is less open to interpretation. The more concrete the requirement, the less the interpretation required by the evaluator, resulting in reduced evaluator bias. Some of the quantifications may be less concrete than others. Some requirements may be constructed in terms of ranges of acceptable performance for differing tactical scenarios. Still, the ranges serve to bound the amount of interpretation required by the evaluator.

F. SUMMARY

In the introduction of this paper two questions are posed. The first asks if factors of the MCCRES evaluation which are subject to evaluator bias can be identified, and the second asks how these factors can be controlled or controlled for. Three areas in which evaluators may bias the MCCRES were identified: senior evaluator influence, other evaluator bias and MPS interpretation. Three techniques were presented to control or control for these sources of bias: evaluator training, evaluator testing and quantification of subjective MPS requirements.

Discussion of the proposed solutions to the problem of evaluator bias did not address the cost to implement the solutions. Before implementing any of the recommendations, a study of benefit and costs of the solutions would be appropriate. A detailed study of the proposed solutions would be likely to point out several methods of implementation for each, possibly revealing still other solutions not addressed in the report.

REFERENCES

1. House, E. R., Evaluating With Validity, Sage Publications, 1980.
2. Levitan, S. A. and G. Wurzburg, Evaluating Federal Social Programs, W. E. Upjohn Institute for Employment Research, 1979.
3. Reiken, H. W., "Action for What? A Critique of Evaluative Research," in Evaluating Action Programs: Readings in Social Action and Education, ed. Carol H. Weiss, Allyn and Bacon, 1972.
4. Stufflebeam, D. L., W. L. Foley, W. J. Gephart, E. G. Guba, R. L. Hammond, H. O. Merriman, and M. M. Provus, Educational Evaluation and Decision Making, F. E. Peacock Publishers, Inc., 1971.
5. Anderson, S. B. and S. Ball, The Profession and Practice of Program Evaluation, Jossey-Bass Inc., Publishers, 1978.
6. Tracey, W. R., Evaluating Training and Development Systems, American Management Association, 1968.
7. Langston, J. H., "OEO Neighborhood Health Centers: Evaluation Case Study," in Social Experiments and Social Program Evaluation, eds. J. G. Abert and M. Kamrass, pp. 107-121, Ballinger, 1974.
8. Tyler, R. W., Basic Principles of Curriculum Instruction, University of Chicago Press, 1950.
9. Drucker, P. F., The Practice of Management, Harper and Brothers, 1954.
10. Schuster, F. E. and A. F. Kindall, "Management by Objectives: Where We Stand--A Survey of the Fortune 500," Human Resource Management, V. 13, No. 1, Spring 1974.
11. Scriven, M., "Goal Free Evaluation," in School Evaluation, ed. E. R. House, McCutchan, 1973.
12. Eisner, E., The Educational Imagination, McMillan, 1979.
13. Wolf, R. L., "The Use of Judicial Evaluation Methods in the Formulation of Educational Policy," Educational Evaluation and Policy Analysis 1, May-June 1974.
14. Barrett, R. S., Performance Rating, Science Research Associates, 1966.
15. Scriven, M., "Objectivity and Subjectivity in Educational Research," in Philosophical Redirection of Educational Research, ed. L. G. Thomas, National Society for the Study of Education, 1972.
16. Cummings, L. L. and D. P. Schwab, Performance in Organizations, Scott, Foresman and Company, 1973.

17. Klimoski, R. J. and M. London, "Role of the Rater in Performance Appraisals," Journal of Applied Psychology, Vol. 59, No. 4, pp. 445-451, 1974.
18. Love, K. G., "Comparison of Peer Assessment Methods: Reliability, Validity, Friendship Bias, and User Reaction," Journal of Applied Psychology, Vol. 66, No. 4, pp. 451-457, 1981.
19. Holzbach, R. L., "Rater Bias in Performance Ratings: Superior, Self- and Peer Ratings," Journal of Applied Psychology, Vol. 63, No. 5, pp. 579-588, 1978.
20. Scott, R. D., "Taking Subjectivity Out of Performance Appraisals," Personnel, pp. 45-49, July-August 1973.
21. Dayal, I., "Some Issues in Performance Appraisals," Personnel Administration, pp. 29-35, January-February 1969.
22. Gallagher, M. C., "More Bias in Performance Evaluation?" Personnel, pp. 35-40, July-August 1978.
23. Bernardin, H. J., "Effects of Rater Training on Leniency and Halo Errors in Student Ratings of Instructors," Journal of Applied Psychology, Vol. 63, No. 3, pp. 301-308, 1978.
24. Wiley, L. and W. Jenkins, "Selecting Competent Raters," Journal of Applied Psychology, Vol. 48, No. 4, pp. 215-217, 1964.
25. Guion, R. M., Personnel Testing, McGraw-Hill, 1965.
26. Marine Corps Order 3501.2, Vol. II, 9 December 1977.

DISTRIBUTION LIST

	<u>No. of Copies</u>
1. Defense Technical Information Center Cameron Station Alexandria, Virginia 22314	2
2. Library, Code 0142 Naval Postgraduate School Monterey, California 93943	2
3. Associate Professor K. J. Euske Code 54Ee Department of Administrative Sciences Naval Postgraduate School Monterey, California 93943	10
4. Colonel Joseph F. Mullane, USMC Code 0309 Marine Corps Representative Naval Postgraduate School Monterey, California 93943	10
5. Commandant of the Marine Corps (Code POR) Headquarters, Marine Corps Washington, D. C. 20380	15
6. Captain George M. Wheeler, USMC 705 Fifth Street Marietta, Ohio 45750	2
7. Office of Research Administration Code 012A Naval Postgraduate School Monterey, California 93943	1
8. Center for Naval Analyses 2000 N. Beauregard Street Alexandria, VA 22311	1

END

FILMED

12-85

DTIC