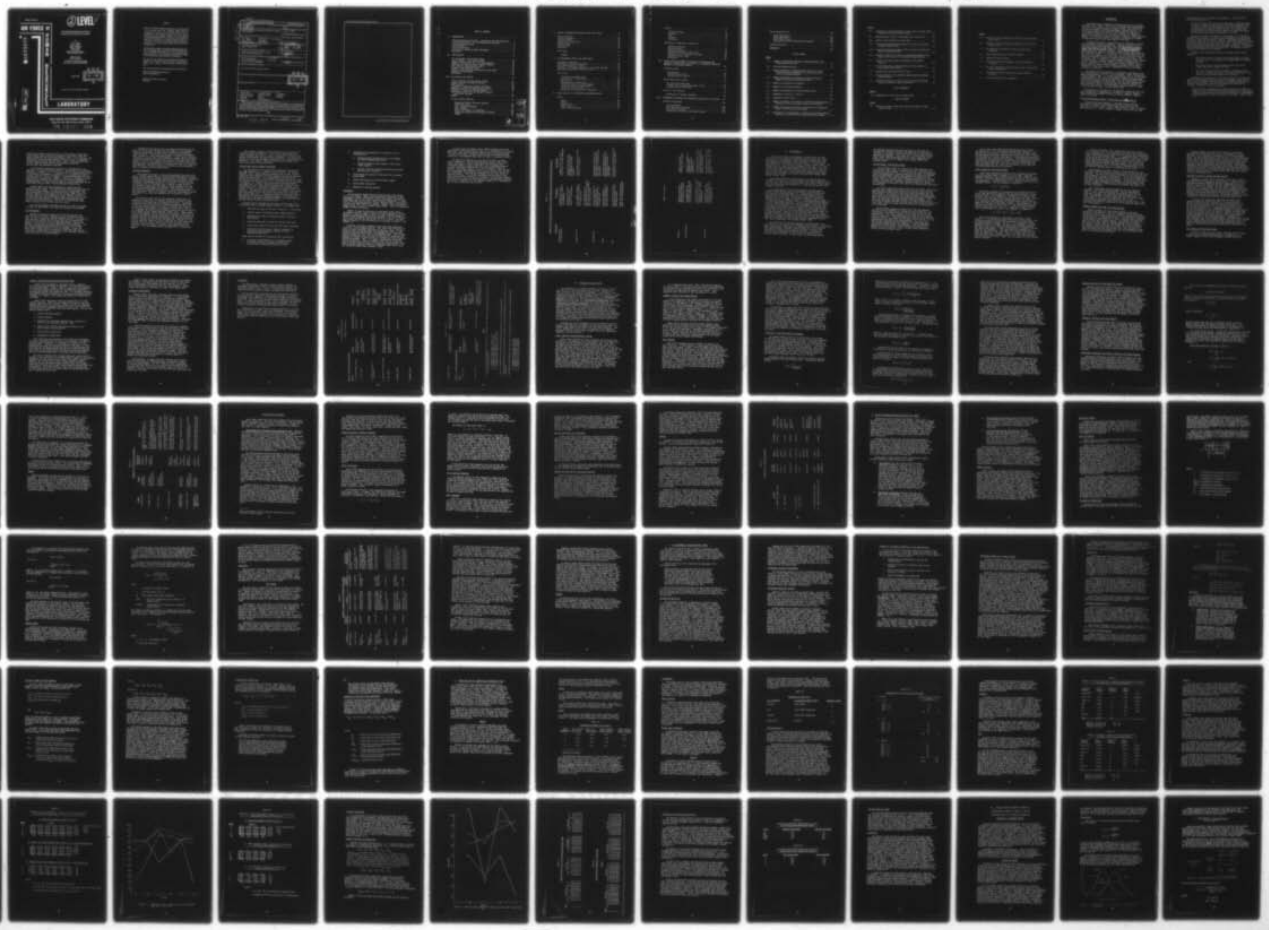
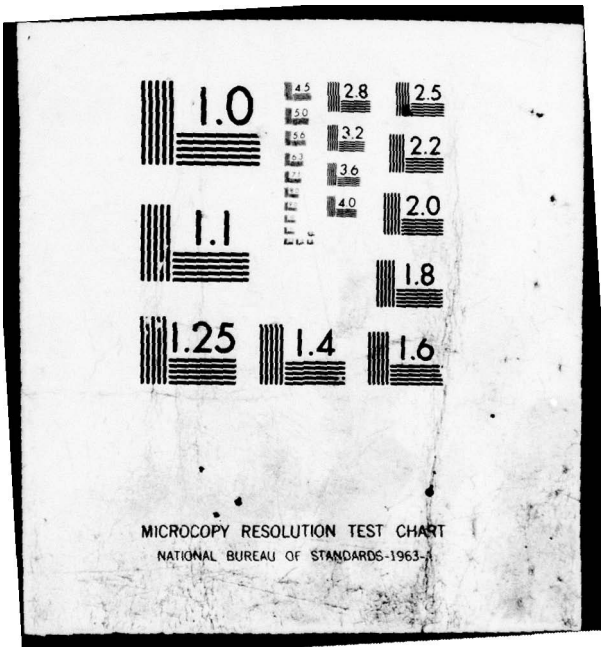
AD-A074 539 APPLIED PSYCHOLOGICAL SERVICES INC WAYNE PA F/G 5/9
CRITERION REFERENCED TESTING: REVIEW, EVALUATION, AND EXTENSION--ETC(U)
AUG 79 A I SIEGEL, L L MUSETTI, P J FEDERMAN F33615-77-C-0046
AFHRL-TR-78-71 NL

UNCLASSIFIED

1 OF 2
AD
A074539




1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

AFHRL-TR-78-71

(2) LEVEL II

AIR FORCE HUMAN RESOURCES LABORATORY

AD A074539

DDC FILE COPY

CRITERION-REFERENCED TESTING: REVIEW, EVALUATION, AND EXTENSION

By

Arthur I. Siegel
Larry L. Musetti
Philip J. Federman
Mark G. Pfeiffer
Joel P. Wiesen
Applied Psychological Services, Inc.
Wayne, Pennsylvania 19087

Philip J. DeLeo
Walter R. Shepperd
TECHNICAL TRAINING DIVISION
Lowry Air Force Base, Colorado 80230

August 1979

DDC RECEIVED OCT 3 1979 B

Approved for public release; distribution unlimited.

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235

79 10 01 054

NOTICE

When U.S. Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

This final report was submitted by Applied Psychological Services, Inc., Wayne, Pennsylvania 19087, under contract F33615-77-C-0046, project 2313, with Technical Training Division, Air Force Human Resources Laboratory (AFSC), Lowry Air Force Base, Colorado 80230. Dr. Philip J. DeLeo (TTT) was the Contract Monitor for the Laboratory.

This report has been reviewed by the Information Office (OI) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public, including foreign nations.

This technical report has been reviewed and is approved for publication.

MARTY R. ROCKWAY, Technical Director
Technical Training Division

RONALD W. TERRY, Colonel, USAF
Commander

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: AFHRL-TR-78-71	2. GOVT ACCESSION NO. / 3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): CRITERION REFERENCED TESTING: REVIEW, EVALUATION, AND EXTENSION	5. TYPE OF REPORT & PERIOD COVERED: Final rept.
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Arthur I. Siegel, Larry L. Musetti, Philip J. Federman, Mark G. Pfeiffer, Joel P. Wiesen, Philip J. DeLeo, Walter R. Shepperd	8. CONTRACT OR GRANT NUMBER(s): F33615-77-C-0046
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Applied Psychological Services, Inc. Wayne, Pennsylvania 19087	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 61102F 2313T403 TT
11. CONTROLLING OFFICE NAME AND ADDRESS: HQ Air Force Human Resources Laboratory (AFSC) Brooks Air Force Base, Texas 78235	12. REPORT DATE: August 1979
	13. NUMBER OF PAGES: 122
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): Technical Training Division Air Force Human Resources Laboratory Lowry Air Force Base, Colorado 80230	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

DDC RECEIVED OCT 3 1979 B

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

checklist predictor	performance checklists	test development
criterion referenced tests	performance evaluation	test reliability
instructional evaluation	rater bias	test validity
item analysis	rating error	
item writing	theory of signal detection	

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The literature relative to criterion referenced test development is reviewed. Rater error in criterion referenced performance evaluation is discussed, and a statistical model for reducing such bias in Air Force applications is presented and experimentally evaluated. The results suggest the utility and applicability of the method in Air Force applications. Needed research into criterion referenced testing in the Air Force is described. The results of a field study into criterion referenced testing in Air Force technical training courses are presented and the implications of the results for Air Force technical training are given.

DD FORM 1 JAN 73 1473

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

031 800

TABLE OF CONTENTS

I. INTRODUCTION 1

Criterion-Referenced Testing: Definitions and Characteristics 1
Criterion-Referenced versus Norm-Referenced Testing 3
Item Writing Differences 4
Use Differences 5
Scoring Differences 6
Instructional Design Feedback Advantages 7
Discussion 8

II. ITEM ANALYSIS 12

Cox and Vargas' Item Validity Index 13
Ivens' Measure of Item Effectiveness 14
Rahmlow, Mathews and Jung's Two Group Approach 15
Hambleton and Gorth's Delayed Posttest Measure 16
Hus's Mastery Discrimination Index 16
Popham's Item Uniqueness Index 17
Crehan's Instructed and Noninstructed Group Index 19
Haladyna's Mixed Method 20
Discussion 21

III. RELIABILITY AND VALIDITY 24

Ivens' Stability and Equivalency Indices 24
Popham's Internal Consistency Measure 25
Unks' Methods 25
Livingston's Traditional Measure Analogues 26
Hambleton and Novick's Replicability Concept 29
Haladyna's Internal Consistency Measure 29
Swaminathan, Hambleton, and Algina's Decision Consistency Approach 29
Summary 33

CR Test Validity Measures 35

Popham and Husek's Construct Approach 36
Ivens' Gain Score 36
Cox's Construct Approach 37
Unks' Concepts 37
Shriver and Foley's Job Relevance 38
Sweezey and Pearlstein's Concurrent Validity 38
Summary 39

ite Section
ff Section
ABILITY CODES
/or SPECIAL
A

IV. MASTERY DETERMINATION/DEFINITION AND TEST LENGTH 41

Delphi Technique 42
Statistical Models 43
Empirical Models 43
Probabilistic Approaches 43
Binomial Model 46
Bayesian Model 47
Discussion 49

Test Length 49

Summary 52

V. CR PERFORMANCE TESTING AND RATER ERROR 53

Performance Checklists 53
Performance Checklist Development 54
Performance Checklist Scoring 54
Examples of Performance Checklists in the Navy and Army 55
Performance Checklists in the Air Force 56

Rater Error 57

Systematic and Random Error 57
General Model of Rater Error 57
Rater Bias 58
Minimization of Rating Error 60
Statistical Models of Rater Behavior 60
Guilford's Model of Rater Behavior 61
Correction for Rater Bias 63
Extension of Guilford's Bias Correction 64

VI. RATER BIAS AND ITS CORRECTION--EXPERIMENTAL STUDY 65

Method 65

Sample 66
Raters 66
Introductory 67
Rater Training 67
Control Over Conditions 67

Results 67

Variance Analysis 68
Leniency 70
Halo 70
Contrast 72
Discussion 72

Estimation of Magnitude of Rater Bias 73

Leniency Correction 73
Halo Correction 73
Contrast Correction 78
Overall Correction for Rater Bias 78
Evaluation of the Correction for Bias 81
Variance Analytic Check 83
Discussion 83

VII. SIGNAL DETECTION THEORETIC APPROACH TO ESTABLISHING THE VALIDITY, DECISION AXIS, AND UTILITY OF A CHECKLIST PREDICTOR EMPLOYED IN A TRAINING CONTEXT 84

Validation Study 84

Calculations 85
Likelihood Ratio (L_X) 87
Interpretation of Results 87

Establishing Cutoff Scores 87

Hit and False Alarm Rates 91
Optimum Value of Likelihood Rater (β Opt) 93
Reliability and Rater Error 94

Utility of the Checklist 98

Advantages of Suggested Approach 98

VIII. PROGRAMMATIC RESEARCH INTO PERFORMANCE CHECKLISTS IN THE USAF. . 100

Checklist Development 100

Item Development 101
Test Objectives and Length 102
Item and Test Scoring 102
Discussion of Checklist Development Studies 103

Rater Characteristics 103

Rater Qualification 104
Rater Experience 105
Rater Point of View 105
Discussion of Rater Characteristics Research 106

Testing Conditions 106

REFERENCES 108

LIST OF TABLES

Table

1.1 Summary of Differences Between Criterion-Referenced and Norm-Referenced Testing 10

2.1 Criterion-Referenced Item 22

3.1 Summary Information for Hypothetical Data on the Joint Classification of Examinees into Two Mastery States on Two Test Administrations 31

3.2 Summary Information for Hypothetical Data for Various Examples of Internal Consistency Measure 32

3.3 Summary of CR Reliability Measures 34

3.4 Summary of CR Validity Measures 40

4.1 Summary of Approaches for Setting Cutoff Scores 50

6.1 Background of Raters for Course A 66

6.2 Components of Rater Bias 68

6.3 Summary of Analysis of Variance of Satisfactory-Unsatisfactory Ratings for Six Teams on Four Traits by Four Raters . . . 71

6.4 Summary of Analysis of Variance of Numerical for Ratings Six Teams of Four Traits by Four Raters 71

6.5 Estimation of the Contribution to Rater Bias for Rater Effects (Leniency), and Rater-Ratee Interaction Effects (Halo) . 75

Table

6.6 Estimation of the Contribution to Rater Error for Rater-Trait Interaction Effects (Contrast Bias) 77

6.7 Original Course A Data in Percentage Score Form Traits . . . 80

6.8 Course A Data Corrected for Rater Bias 80

6.9 Intraclass Correlation Among Raters by Trait for Corrected and Uncorrected Scores 82

6.10 Intraclass Correlations Among Traits for Halo Corrected and Halo Uncorrected Ratings 82

6.11 Summary of Analysis of Variance of Numerical Ratings of Six Teams on Four Traits by Four Raters (Adjusted Data) . . .

7.1 Sequence with which Raters are Exposed to the Payoff Matrices 89

7.2 Data Obtained From a Single Observer in Five Sessions in Which Payoffs Were Varied 91

7.3 Optimum B for Conditions Given in Text 93

7.4 Likelihood Ratios of 10 Observers Employing Five Different Payoff Matrices 94

7.5 Differences Between β' and β' Opt for 10 Raters Operating Under Five Different Payoff Conditions 95

LIST OF EXHIBITS

Exhibit

6.1 Fragmentation of Part One into Task Items 69

LIST OF FIGURES

Figure

6.1 Numerical rating by means over traits and teams by raters; leniency bias 74

Figure

6.2 Numerical rating of means over traits by teams by raters: halo effect 76

6.3 Numerical rating of means over teams by rater and trait: contract bias 79

7.1 Distributions of checklist scores for masters and nonmasters 85

7.2 Cross-classification of validation sample 86

7.3 Symmetrical payoff matrix 88

7.4 Five payoff matrices designed to vary the observer's criterion 90

7.5 ROC curve obtained by varying payoffs rather than presentation probabilities 92

7.6 Relationship between obtained and optimal decision criteria 96

7.7 Cross classification of candidates 97

I. INTRODUCTION

The present report presents a detailed discussion of criterion-referenced (CR) testing. The report is organized into eight chapters. The first five chapters discuss a number of major topics within CR testing and attempt to describe the general state-of-the-art. The next chapter reports on an empirical examination into the state of CR testing within the U.S. Air Force. Chapter VII suggests an approach to validation of CR testing which is based on Signal Detection Theory. And, the final chapter is devoted to developing a program of research tailored to Air Force needs in this area.

To examine the overall state-of-the-art of CR testing, an extensive literature review was undertaken. Computer searches of two information banks were completed. Both the Psychological Abstracts Search and Retrieval System (PASAR) and the Defense Documentation Center (DDC) data banks were searched. These computer searches were conducted to identify literature relevant to the major aspects of the literature review; (1) CR testing, (2) criterion checklists, (3) models of rater behavior, (4) rater bias, and (5) correction for bias. Each search covered a period of 10 years ranging from January, 1967 to January, 1977. In addition to the computer searches, normal reading of the literature uncovered a number of relevant references and a body of literature. In all, the computer searches together with the manual search provided a rather extensive coverage of the relevant literature in the various areas of psychometric practice and theory relative to CR testing.

The results of the two empirical investigations performed by the Applied Psychological Services are also reported. These investigations were conducted through: (1) a case study approach and (2) a field experimental approach. The case and experimental studies were undertaken to examine the state of criterion checklists and CR performance testing in Air Force resident training courses. The case study approach was used to discover modes of criterion checklist usage at selected Air Force technical schools. The experimental study was based on a model criterion checklist used in an Air Force course.

On the basis of the results of the empirical studies and the literature indications, recommendations are made and a general research plan presented with respect to future criterion checklist research in the U.S. Air Force.

Criterion-Referenced Testing: Definitions and Characteristics

Criterion-referenced testing, as the name implies, measures performance in terms of demonstrated proficiency--usually as the result of a course of study. Glaser (1963) pointed out that, in achievement testing, it is vital to compare a student's test score with some

criterion of mastery (a standard of performance). Glaser and Nitko (1971) defined mastery as:

> The term "mastery" means that an examinee makes a sufficient number of correct responses on the sample of test items presented to him in order to support the generalization (from this sample of items to the domain or universe of items implied by an instructional objective) that he has attained the desired, pre-specified degree of proficiency with respect to the domain. (p. 641)

One way of looking at CR testing is to consider it as a method of interpreting test scores. CR tests yield measurements that have meaning in terms of specified performance standards. Performance standards are specified by defining "a class, or domain of tasks that should be performed by the individual" (Glaser and Nitko, 1973, p.65). Since the CR test is designed to provide information about performance standards, the standards must be established prior to test construction. The student's status is assessed with respect to the standards.

Gronlund (1973) described CR tests as including several characteristics:

1. CR testing requires a defined and delimited domain of tasks, where the focus is on mastery of a limited number of learning outcomes.

2. CR testing requires that performance objectives must be clearly specified in behavioral terms.

3. The score should describe the student's performance on the task (e.g., student can perform 70 percent of the steps in the technical orders without instructor assistance).

Although most definitions of a CR test are similar in content and emphasis, the definition offered by Harris and Steward (1971) (cited in Atkin, 1974) is somewhat different because it emphasizes sampling--an important concept when a wide variety of behavioral objectives is involved. According to Harris and Steward:

> A pure criterion-referenced test is one consisting of a sample of production tasks drawn from a well-defined population of performances, a sample that may be used to estimate the proportion of performances in that population at which the student can succeed. (Atkin, 1974; p. 4)

Rather than concentrating on performance standards, other definitions of CR testing call attention to the different types of CR tests.

According to Popham and Husek (1969) there are essentially two types of CR tests. In the first type, items relate directly to the criterion and the test is homogeneous so that everyone achieving the same score on the test obtained it in the same way (i.e., a specific score indicates a particular response pattern). This test type is taken from Guttman's (1944) concept of reproducibility. The second type of CR test contains samples of items from a population of items that are directly related to the criterion. Popham and Husek found the former type of CR test to be ideal, but its use is restricted primarily to formal, well structured areas. The latter type of CR test is more typical, but contains an element of ambiguity. If a number of items is missed, one can not tell which ones from the score alone. Accordingly, without a more detailed analysis one can not know specifically what the student can and can not perform.

Gronlund (1973) differentiated between CR "mastery" tests and CR "developmental" tests. In the mastery case, the students are expected to be able to achieve some level of accuracy (say, 80% correct). This is due to the elementary level of the material. Developmental areas, on the other hand, include complex behaviors (e.g., "applies concepts and principles to new situations," and "writes a creative short story") which may never be fully mastered. Due to little score variance, Grounlund suggested that there are no acceptable correlational statistics available for use in measuring the reliability or validity of mastery level CR tests. However, the standard correlation statistics can be used with the developmental level CR tests. Gronlund, therefore, indicated that more care should be taken in the development and construction of CR mastery tests.

The modal definition of CR testing seems to emphasize interpretation of test performance in terms of demonstrated behavior. The CR test is "based on set of specific learned objectives stated in terms of behavioral changes to be expected in the examinee as a result of instruction toward these objectives." (Roundabush & Green, 1972)." In short, the emphasis is on what the examinee can do; the skills he can display.

Criterion-Referenced versus Norm-Referenced Testing

The CR method of interpreting test scores is contrasted with the more traditional norm-referenced (NR) interpretation. Unlike NR testing, CR testing is not concerned with how well the examinee compares with others. NR tests indicate proficiency relative to a norm group. For the CR tests, the focus is on what the examinee can do, not on how the examinee compares with others.

Distinctions between CR testing and NR testing have been discussed by a wide number of investigatiors (Anastasi, 1976; Livingston, 1972; Ivens, 1970; Popham & Husek, 1969; Simon, 1969; Hambleton & Gorth, 1971; Glaser, 1971; Glaser & Nitko, 1971; Gronlund, 1973; Alvord & Buttingham, 1974).

Garvin (1971), for example, distinguished CR and NR tests with regard to the subject matter of testing. He suggested that CR and NR testing are applicable to different sets of subject matter. This distinction is based on the tasks involved. Some tasks demand a very high level of performance each time the task is performed (e.g., landing an airliner). For such tasks, CR testing is appropriate. Other tasks inherently have more variability in performance, but yet some task-related criterion is specifiable (e.g., house painting, balancing a checkbook). For these types of tasks either NR or CR tests are appropriate. Finally, there are tasks for which no objective criteria are inherent in the subject matter. In these cases, it is only possible to create arbitrary criteria, such as, the ability to do 10 pushups within one minute. For this latter type of task, NR testing is appropriate. Accordingly, Garvin sees CR testing as more appropriate: (1) when performance standards are objective and specifiable, and (2) when the task requires completely successful performance each and every time it is performed.

Glaser (1971) suggested social studies as an example of a course of study in which CR testing is not appropriate because no meaningful performance criteria can be developed. He identified tasks on jobs involving public safety and jobs where critical financial and ethical standards are involved as examples where CR testting could be used. CR testing was also identified to be of value in licensing individuals for various professions, (e.g., doctors, lawyers, pilots). Glaser (1971) also suggested the CR test for use in controlling entry into successive courses of study that follow previous instructional sequences, as in reading and the physical and biological sciences.

Item Writing Differences

There are also those who hold that CR and the NR tests also rest on different item writing approaches. Such a difference, if true, seems of equal, if not greater, saliency than the behavioral emphasis and interpretive differences mentioned above.

The undesirability of the traditional approaches to item writing when constructing CR tests was discussed by Bormuth (1970), Gagne (1969), Gronlund (1973), Hambleton & Gorth (1971), Ivens (1970), and Popham & Husek (1969). In writing test items for an NR test, where

maximizing variance of test scores is a prime concern, item writers often prepare items using novel language or with options that are equally appealing (for the multiple-choice item). For CR testing, such techniques are not advocated because spurious factors or nonlearned material may be written into the test. That is, items written for a CR test should directly measure the behavioral objectives that the test was designed to evaluate.

In CR testing, operational definition of the behaviors to be measured is critical to the item writer. Specificity in the criterion objective is important if it is to be used as the standard against which an individual's achievement is compared. It is acceptable to test a student's ability to use a voltmeter in troubleshooting when the criterion objective so specifies. However, it would be confounding if the objective only indicated the ability to use auxiliary equipment in troubleshooting. In short, writing items or constructing tasks for CR tests must directly follow from the behavioral objective(s).

Several models have been suggested for constructing and selecting CR test items. Klein and Kosecoff (1973) outlined three of the more popular methods: (1) employ a group of measurement and curriculum experts to develop and decide on test items to use, (2) develop a matrix of tasks (behaviors) to be assessed for each behavioral objective and systematically sample from this matrix, and (3) construct items and use a computerized item generating technique for covering objectives. The first of these methods seems highly practical but of minimal empirical value.

The item development difference question is further discussed in Chapter II when formal item analytic procedures are presented.

Use Differences

Another distinguishing feature of CR testing concerns the specific uses of CR tests. Whereas NR tests are held to be practical for use in selecting the best (highest scorers) in a group, CR tests are more suitable for selecting those who can demonstrate mastery of a particular skill or those who possess a certain competence. NR tests do not address themselves to the questions: (1) What learning outcomes were measured by the test? and (2) What are the achievement levels of a class of students? Glaser and Nitko (1971) claimed that NR testing has limitations when the instructional system seeks to be adaptive to the individual student because the NR test is seldom diagnostic.

Although NR tests can be used to evaluate an instructional program, as can CR tests, NR tests are less useful for this purpose. NR tests are designed to measure and to maximize discriminating power (variance). The CR test may be used to evaluate instruction more directly as in the following example: if test results indicate variations in passing and failing among the students who have recently received training on a relevant task, then the instruction was inadequate. If all of the students can perform a task after instruction that they could not perform prior to instruction, then the instruction was effective. The ideal spread in CR testing is zero--indicating that all students have attained mastery.

Scoring Differences

The scoring of CR tests is also distinguishable from that of NR tests. Neither percentiles nor standard scores characterize CR tests. Normative scores would contradict the meaning of CR testing. Generally, CR test scoring involves establishment of minimum essential scores. Standards of performance can be set in terms of: (1) the precision with which a task is performed (e.g., aligning a meter to within five degrees of zero), (2) the number of errors allowed (e.g., performing the minimum number of procedural steps required), (3) the time required to complete the task (e.g., locate the malfunction in the electrical circuit in 15 minutes), and/or (4) the condition of the end product (e.g., does the equipment work following an internal repair).

On the other hand, some CR test proponents contend that although scoring a CR test is feasible and may be desirable in some situations, different levels of mastery do not exist. Roundabush and Green (1972) stated that the assumption underlying any CR test is that the student either has or has not mastered the objective. No continuum of achievement is assumed as with NR tests. That is, it is assumed that mastery level students possess one level of ability and nonmastery level students another (lower) level. These assumptions, taken together, imply that the distribution of scores on a CR test will be bimodal with the lower mode interpreted as the mean for students who fail to perform the behaviors, and the upper mode the mean for students who can perform the behaviors. With this, an all-or-none, pass-fail score is seen as sufficient for scoring CR tests, (Emrick, 1971; Hsu, 1971; Unks, 1971; Shriver and Foley, 1974). According to this thinking, a "mastery" score should answer the question: "Can the student perform the behavior?" in a yes-or-no manner.

Other authors, however, maintain that pass-fail scoring loses too much valuable information. In particular, Popham and Husek (1969) suggested that there is useful information in failure or nonmastery scores beyond the indication of nonmastery. If alternative courses of action are available, reporting of this information may be beneficial. For example, remedial efforts could be based on the nature and extent of the errors made.

Instructional Design Feedback Advantages

The implications of CR testing for instructional design have also been pointed out in a number of prior reports. For example, Shoemaker (1971) claimed that with sequential CR testing, instructional decisions regarding individual students and the instructional program can be maximized. His recommendation was based on pre and posttraining tests. The pretraining test would consist of items from the next instructional unit (i.e., following the one the student is currently in). Items in the posttraining test would measure achievement over the global objectives which were covered in past instructional units. According to Shoemaker, the advantage of the pretest is that it can be used to determine the pace for future instruction. The pace could be increased if the student receives more than the minimum score on the skills required for entry into the next instructional unit. A delayed posttraining test could also be designed for use with the CR measuring scheme. The delayed posttraining test is designed to measure retention of instructional objectives covered in past instructional units.

Gronlund (1973) indicated that CR tests could be used in the instructional plan to improve student learning. His procedure was:

a. administer CR tests at the end of each unit of instruction

b. analyze results to determine steps student mastered

c. examine failed items to determine student's learning deficiencies

d. prescribe additional training for the deficient areas

e. retest with a parallel form of the test after retraining

f. use results from the test to improve subsequent instruction by way of methods, materials, and/or sequencing modification.

Other uses of CR tests for instruction and training are:

- an analytic and diagnostic tool to determine when to advance students to the next subject matter area and when to provide additional training

- pretesting at the beginning of a course or unit of instruction to:
  a. determine which prerequisite skills are needed for the forthcoming instruction
  b. determine where to place student in the instructional sequence
  c. provide a base for measuring learning gain during the course of instruction
- at the end of the course of instruction for assigning course grades
- predict job success for the individual student
- individualize instruction
- determine if learning occurred

Discussion

The differences between NR and CR testing seem to be neither trivial nor surface. They include differences in approach to test development, content differences, scoring differences, interpretive differences and, in some cases, administrative differences. In some cases, the differences may also involve "adverse impact", differential validity, and equal opportunity nuances. However, the literature has spoken little, if at all, to these latter points. Table 1.1 attempts to summarize these differences.

Whether CR tests and NR tests are considered points along a continuum or distinct types of tests, the scores yielded by, and the purpose of, these two types of tests are often different enough to require rather different statistical treatment. The crux of the matter is score variability and the lack of it in CR tests. The more technical aspects of this distinction are discussed in Chapter II and Chapter III.

Typically, CR tests are mastery tests with large proportions, if not all, of the students passing. This lowers the variability in scores for both individual items and the test as a whole. Further, CR tests often are (purposefully) very homogeneous, with only one objective being tested by a pool of homogeneous items. This further reduces variability. Now, the classic item analytic techniques and measures of reliability and validity are largely or entirely based on correlation methods. And, as is known, correlational indices depend on variability. Thus, for example, under the low variability conditions which usually exist in a CR testing situation even a test that is highly stable and internally consistent might yield a very low reliability coefficient (Anastasi, 1976).

A classic paper by Popham & Husek (1969) presented the earlier approaches to the problem posed by the lack of variability of CR scores. Their work shows how the classical measures may be ineffective when CR tests are involved. Popham & Husek called for new and different statistics, less dependent upon score variance, to evaluate CR tests.

The search for measures of the worth of CR tests has taken several approaches: applications and adaptations of classic techniques, analogues to classic techniques, new techniques which intuitively appear to measure what the classic techniques measure, and other techniques based on one or more assumptions about CR testing. CR testing, however, is yet in its infancy. Even the most basic techniques and procedures for evaluating CR tests are not yet fully developed. Those that are developed are not fully tested in empirical settings. The various procedures and statistical techniques that have been proposed are reviewed in the following chapters. Empirical data relevant to each technique are reviewed when available. The review is organized into two separate chapters--(1) item analysis and (2) test reliability and validity.

Table 1.1

Summary of Differences Between Criteion-Referenced and Norm-Referenced Testing

Difference	CR Testing	NR Testing
Interpretation of Test Scores	compares test score with absolute standard of performance	compares test score with the scores of other examinees
	indicates the extent the examinee is a master of performance	indicates how well the examinee has performed as compared to other students
Subject Matter	tests subject areas in which defined standards of success exist	tests subject areas in which success is undefined
	more appropriate for testing tasks which require high levels of performance each time performed	
Item Writing	directly considers the content domain and behavioral objectives of instruction	may consider spurious and other factors unrelated to the content or objectives of instruction
Focus	concerned primarily with the individual not individual differences	concerned primarily with individual differences and discrimination between individuals
Use	most useful for making absolute decisions on an individual basis	most useful for making relative decisions on a group basis
	better for evaluating the effect of training programs	

Table 1.1 (cont.)

Scoring	based on objective performance standards and pass-fail scoring	based on norm and standard score tables
Score Variance	no attempt made to increase score variance, via test construction	typically attempts to increase score variance via test construction
	generally produces scores with little or no score variability and skewed distribution	generally produces scores with adequate score variability and approximately normal distribution
Statistics	traditional test construction indexes and correlational statistics are inappropriate	traditional item analytic, test reliability, and test validity indexes are appropriate

II. ITEM ANALYSIS

Item analysis refers to procedures for determining item characteristics for the purpose of selecting "good", useful test items. Traditionally, with NR tests, three important characteristics have been employed to determine an item's usefulness: (1) correlation with other items, (2) difficulty, and (3) validity (Guion, 1965). And traditionally, a "good" NR item is defined as one with: (1) a high average item intercorrelation coefficient, (2) an optimum item difficulty level, and (3) a high validity coefficient (i.e., high item-total score correlation). Such traditionally accepted item characteristics are not useful for CR test item selections (Gagne, 1969; Popham & Husek, 1969).

Due to low variability within scores, a CR item may show low internal consistency (e.g., low correlation with other items) and low validity (e.g., low correlation with total score) while actually possessing high consistency and validity (i.e., high correlation with some external criterion).

The interpretation of item discrimination indices presents another problem when CR tests are considered. With NR tests, only positively discriminating items (i.e., items answered correctly more frequently by the high total test scoring students) are considered to be "good" items. For CR tests, however, a different viewpoint exists. CR items are "good" if (and some might say only if) they have content validity; that is, if they measure the content of a specific behavioral objective. Item discrimination indices, therefore, are used differently in CR test development. Positively discriminating items might be used to locate areas requiring additional or modified instruction so that all students can be brought to the mastery level--a goal of the instructional program designed under the CR testing concept. Negatively discriminating items (i.e., items answered correctly more often by the low total test scoring students) may serve to identify a need for revision either in the items or in the focus of instruction and instructional material (Hambleton & Gorth, 1971; Gorth & Hambleton, 1972; Popham & Husek, 1969).

The traditional correlational methods are inappropriate to CR tests because of the lack of score variance within CR tests. Accordingly, other item analytic techniques, less dependent upon score variance, have been developed. The next sections present a review of the item analytic techniques developed for CR tests. The review describes several techniques for calculating CR item difficulty, validity, and reliability. In most cases, the assumptions behind

and problems associated with each technique are discussed and the empirical evidence available with regard to the item analytic technique is presented. Eight articles and their respective item measures are discussed. The authors are: Cox and Vargas (1966), Ivens (1970), Rahmlow, Mathews and Jung (1970), Hambleton and Gorth (1971), Hsu (1971), Popham (1971), Crehan (1974), and Haladyna (1974).

Cox and Vargas' Item Validity Index

Cox and Vargas (1966) (as cited by Cox, 1971) developed a CR item validity index based on a comparison of test responses prior to training and after training. This item validity index is computed by taking the percentage of students who passed an item on the posttest minus the percentage of those same students who passed the item on the pretest. The higher the difference, the greater the item validity. Note here Cox and Vargas' assumption that prior to training students are nonmasters, and after training students are masters. This assumption is violated to the extent that learning and instruction are unrelated.

Cox (1971) summarized another study by Cox and Vargas (1966) which compared their posttest minus pretest item validity index with the more traditional upper minus lower group index (calculated only for the posttest). The results of this study indicated that items found to be valid using the prepost test paradigm failed to discriminate between the upper and lower total posttest scoring students. From this, Cox and Vargas concluded that the traditional method was sufficiently different from their prepost test index to warrant the use of their index, especially where score variability is not the concern--as with CR tests.

The Cox-Vargas argument is based on the logic that because CR tests are not concerned with increasing score variability, then item analytic techniques which are based on score variability are inappropriate. This logic, however, is not based upon the cornerstone of CR testing. The purpose behind CR tests is not to decrease score variability, but rather to learn what a person can do, i.e., to differentiate the masters from the nonmasters. As such, a CR item validity technique should attempt to find those items which discriminate between masters and nonmasters. (By inappropriate selection of cutoff scores, the definition of masters and nonmasters could be fallacious. Such a situation would negate the utility of the Cox-Vargas index.)

The Cox and Vargas index defines mastery according to training received (i.e., if in the pretraining condition, nonmaster; if in the posttraining condition, master) while the upper and lower index defines mastery according to relative total test score (e.g., if score is above 50% of all scores, then master). The two mastery definitions are different (but not mutually exclusive). The possibility exists that a person will be classified as a master under one definition but not under the other definition. With such a criterion difference, one might anticipate that the indices will disagree on which items discriminated between masters and nonmasters.

Ivens' Measure of Item Effectiveness

Ivens (1970) developed two measures which attempted to evaluate the overall effectiveness ("goodness") of a CR item. These two measures combined an index of item reliability with an index of item validity. Three test administratiions are required for calculation of the two measures: a preinstruction test, a postinstruction test and a subsequent retest. The first measure is defined as:

$$f_1 = (1 - P_{ab})(P_{bc})$$

where P_{ab} is defined as the proportion of students whose score on the item remained the same over the two administrations, pre and postinstruction. This component is considered as an index of item validity: the lower the proportion, the more valid the item. That is, a valid item is one whose score changes from the pre to posttest. P_{bc} is defined as the proportion of students whose scores on an item remain the same from the posttest to retest. This component is considered as an index of item reliabilitv: the higher the proportion, the more stable the item. Thus, $f_1 = 1$ would be a perfect CR item. Iven's second measure of overall item "goodness" is indexed as:

$$f_2 = (P_b - P_a)(1 - |P_c - P_b|)$$

The $(P_b - P_a)$ component is considered as the index of item validity. It is the Cox and Vargas definition of item validity: the difference between the proportion of students passing the item on the posttest and the pretest. The $(|P_c - P_b|)$ component is defined somewhat similarly. P_c equals the proportion of students passing the item on the retest. P_b equals the proportion of students passing the item on the posttest. Ivens considered this component as an index of item reliability: the higher the absolute difference, the lower the stability of the item. That is, a stable item is one on which the proportion of students passing that item remains the same from the post to the retest. Again, as f_2 increases to unity the item quality increases.

Like Cox and Vargas, Ivens assumed that mastery and instruction go hand in hand. This assumption causes the item validity index to be confounded by instructional quality. Besides this confounding, Ivens' first item validity index, f_1, is apparently biased. According to the definitions of item validity and mastery, a valid item is not one for which the response just changes, but one on which the response specifically changes from fail to pass. The f_1 index considers both the fail-to-pass response change and the pass-to-fail response change as valid. The f_1 index is thereby biased, as the pass-to-fail response change is, by definition, not valid.

Some problems also exist relative to the multiplicative combination of the two components (reliability and validity) of the Ivens equations. (We ignore the question of whether or not combining the two psychometric concepts is meaningful.) In the Ivens conception, because of the multiplications involved, a low value of either factor will reduce the index value considerably. The multiplicative combination assumes independence for reliability and validity--a probblematic assumption. Other combinatorial methods seem equally, if not more, defensible.

Ivens (1970) empirically evaluated his two item "goodness" measures. Two CR tests were developed. One was considered a "good" CR test, while the other was considered a "poor" CR test. Data were collected, and item analysis was performed using Ivens' item measures. Results showed that the item indices tended to be higher in the "good" than the "poor" test. Item reliability was generally not different across the two tests. Ivens concluded that the exact properties of the two measures require further investigation.

Rahmlow, Mathews and Jung's Two Group Approach

Rahmlow, Mathews, and Jung (1970) presented a two group approach to item analysis. With this approach, one group receives the CR test prior to any instruction, while the second group receives instruction prior to taking the CR test. The item index is computed as the change in the proportion of correct responses from the noninstruction to postinstruction groups: the greater the change, the more valid the item. However, Rahmlow, Mathews, and Jung noted that such a procedure does not show whether the item was mastered. For example, a positive change score could result with only 5% of the postinstruction group passing that item. They suggested, therefore, that along with the change score, the item difficulty be computed for the postinstruction group.

Rahmlow, Mathews, and Jung (1970) conducted a study which compared three item analytic statistics. Two groups of students were used: (1) a noninstruction group and (2) a postinstruction group. The three item analytic statistics were: (1) the traditional point biserial correlation coefficient, (2) an item difficulty index, and (3) their less traditional non to postinstruction change score. The point biserial coefficients and item difficulty indices were computed on the postinstruction group's test scores. These authors indicated as the result of visual examination of the data, that high ranking point biserial coefficients did not agree with item difficulty indices or change scores. The high ranking item difficulty indices did agree with change scores.

Hambleton and Gorth's Delayed Posttest Measure

Hambleton and Gorth (1971) presented an item index slightly modified from that presented by Cox and Vargas. Instead of the pre-posttesting paradigm, Hambleton and Gorth suggested the use of a pretest, along with a "delayed posttest" (administered one month after the end of instruction). The item index is computed as the proportion of students passing the item on the delayed posttest minus the proportion of students passing the item on the pretest. Again, mastery is defined by instruction and hence again the measure is confounded by instructional quality. The delayed posttest is further confounded by forgetting and learning during the delay period.

In an empirical examination, Hambleton and Gorth (1971) compared three item validity measures: (1) a traditional biserial correlation (based solely on a postinstruction test), (2) the Cox and Vargas measure, and (3) their modification of the Cox and Vargas measure (i.e., a delayed postinstruction test). The results indicated little relationship between the traditional biserial correlational indices and the less standard Cox-Vargas type pre-post (or delayed post) test indices. There was a strong relationship between the Cox-Vargas measure and their delayed post-test modification of the Cox-Vargas measure. Hambleton and Gorth indicated that item selection based on the three indices yielded sets of items, some of which differed widely from each other. The widest difference among the pools of items selected occurred between the traditional biserial and the Cox-Vargas item validity indices.

Hsu's Mastery Discrimination Index

Hsu (1971), along with most authors, defined a valid CR item as one that discriminates between masters and nonmasters. If masters respond correctly while nonmasters respond incorrectly,

then the CR item is valid. As noted, the problem becomes one of determining mastery. Typically, mastery is determined by completion of instruction: a master is one who was instructed. Hsu noted, however, that this definition confounds the measurement of item validity with quality of instruction. Hsu, accordingly, presented two item measures which did not rely on the assumption that instruction produces mastery.

Hsu's (1971) measures are based on a different definition of mastery. Mastery to Hsu is determined by an established cutoff score. A similar concept was previously introduced by Siegel, Schultz, Fischl, and Lanterman (1968), who called their cutoff score an "absolute criterion." With Hsu's definition, mastery is defined without regard to instruction. Hsu, thereby, suggested that an item validity index can be calculated either by: (1) a difference index (Dp): the proportion of masters who respond correctly minus the proportion of nonmasters who respond correctly or, (2) a phi coefficient (d) using the categories: master-correct response, master-incorrect response.

Hsu suggested that if no variability exists within the item scores (e.g., everyone passed) or mastery group (e.g., everyone a master), then either a point biserial coefficient or Dp could be used.

Hsu (1971) examined the correlations between three item validity indices: (1) his Dp index, (2) phi coefficient and, (3) the traditional point biserial coefficient. Using CR test results, Pearson product-moment correlations were calculated between these three indices for pretrained and posttrained students. These conditions produced two distinct test score distributions: (1) a more normal heterogeneous distribution of scores, and (2) a more skewed homogeneous distribution of scores. The results indicated that all three indices were in considerable agreement. However, there was more agreement when the students had a normal, heterogeneous range of scores than when the range was more homogeneous and more skewed. Also, items which discriminated highly among students within a given distribution of test scores were not necessarily the same items which discriminated well with a group of students with a different distribution of abilities. Such findings speak poorly for the various indices. They are evidently not distribution free. If different measures of the same thing do not agree, a problem exists. On the other hand, if one holds that the various measures are measuring different things, why compare them in the first place? However, when the same index was calculated based on two different groups with the same distribution of test scores, there was a large degree of agreement.

Popham's Item Uniqueness Index

In an attempt to measure item uniqueness, Popham (1971) develped a procedure for identifying single items which are not similar to the items comprising the test as a whole. The procedure

rests on developing an overall response pattern for the test which may be compared with the response pattern for a single item. The procedure classifies a student's response to an item on both a pre and posttest into one of four possible categories: C-C, C-I, I-C, or I-I where C = correct and I = incorrect. All items are categorized and the frequency within each category is determined. The median frequency for each category is found across items. These data, taken as prototypic of all items measuring the CR test objective, are used as theoretical frequencies. The observed frequencies are found for each item and a chi-square goodness of fit test is performed. The chi-square analysis is performed for each item. A significant chi-square indicates that the item under consideration does not resemble the theoretical, prototypical item response pattern.

Popham (1971) applied his chi-square approach to item internal consistency and found the results to have face validity. The chi-square index identified those items Popham thought (based on inspection of the data) to be aberrant. Such evidence, however, should be taken with caution. Further research should compare Popham's chi-square item index with other item internal consistency indexes, (e.g., item correlations with other items or total test score given different test score variability conditions.) The pattern analytic approach possesses considerable appeal. However, other indices of profile similarity are available and should be tried.

In addition, Popham (1971) explored the use of pre and posttest item response patterns to examine items on a CR test. Popham reasoned that if instruction improved learning (e.g., produced mastery), then a fail-pass response pattern would reflect this improvement while a pass-fail pattern would not. Thereby, he reasoned that a negative relationship should exist between the percentage of fail-pass responses and percentage of pass-fail responses to an item. Popham performed such an analysis on a limited set of data. He ranked the items within each response category and obtained the intercorrelation. The results showed no substantial negative correlation.

If Popham had obtained significant negative correlations, then how would one interpret such results? Would Popham conclude that overall the items on the CR test are "good"? According to Popham's reasoning, such a conclusion might be incorrect. A negative correlation could indicate that the items have a high percentage of pass-fail responses coupled with a low percentage of fail-pass responses. Such items, however, by Popham's reasoning are "poor". Accordingly, the test would contain some "poor" or indifferent items. Also, because correlation does not consider the absolute position of variables, a negative correlation could occur as a result of items with low percentages in the fail-pass category coupled with even lower percentages in the pass-fail category. Is an item with 5% of the responses as fail-pass and 1% as pass-fail "good"? Such a result could occur if 94% of the responses were within the pass-pass and fail-fail categories.

Crehan's Instructed and Noninstructed Group Index

Crehan (1974) presented an item measure which is based on data from two separate groups of students: an instructed group and an uninstruted group. Crehan's index is computed as the proportion of instructed students who passed an item minus the proportion of uninstructed students who passed that item. This approach eliminates the repeated measure confounding of the single group prepost testing paradigm. It, however, also assumes that quality of instruction is a constant.

Crehan (1974) conducted an empirical evaluation of six item analytic measures. Each of the six measures was used to rank and thereby select items for a CR test. Hence, six pools of items were selected and used to represent separate CR tests. The reliability and validity of the resulting tests were then compared. The six item analytic measures were:

1. The Cox and Vargas measure
2. Crehan's measure
3. Proportion of consistent responses (e.g., pass-pass or fail-fail) on equivalent, parallel items
4. Teacher rating ("which item would you choose if you were to give a one-item test?")
5. Point biserial coefficient
6. Assignment of random ranks

Test reliability was estimated as the proportion of agreement in overall test grades (pass or fail) on two parallel test taken by the same group of students. An agreement, for example, was two passing grades. Test validity was estimated in two ways: (1) the proportion of students who passed the test in instructed group plus the proportion of students who failed the test in the uninstructed group, and (2) the point biserial correlation between the numerical test score and group membership (instructed versus uninstructed).

The results indicated that although there were significant differences in the reliability of the tests based on these six item analytic techniques; no clear pattern of one "best" item analytic measure emerged. For the two test validity indices, however, Crehan's index and the Cox and Vargas measure produced tests which showed significant and consistently higher test validities over the other four item measures. With these results, Crehan (1974) indicated that his item index and the Cox and Vargas index are the preferred methods of item analyzing CR tests.

Crehan's (1974) results are possibly confounded by the measurement methods. Conceptually and operationally, both his item index and Cox and Vargas' item measure are more identical to the two test validity measures than the other four item measures. Accordingly, the question of tautology in Crehan's test remains open.

Haladyna's Mixed Method

Haladyna (1974) presented a two group design for item assessment. He considered three samples of students: (1) preinstructed students who were at nonmastery, (2) postinstructed students who were at mastery, and (3) a combination of these two groups. Mastery level was defined by an arbitrary cutoff score and hence not completely dependent on instruction. That is, an instructed student was not necessarily considered a master and a noninstructed student was not necessarily considered a nonmaster. For an item measure, Haladyna maintained that the traditional point biserial correlational coefficient is an acceptable CR item analytic technique, if enough score variance exists. To obtain score variance, he hypothesized that the problem posed by a lack of variability with CR scores could be overcome if mastery and nonmastery students are considered together. Thus, Haladyna suggested the use of point biserial correlation, calculated on a combined sample of mastery and nonmastery students, as a CR item index.

To test the adequacy of point biserial correlation, Haladyna (1974) compared the correlation with an item difficulty difference (D) measure. This D item measure computed the proportion of postinstructed mastery students who answered the item correctly minus the proportion of preinstructed nonmastery students who answered the item correctly. Point biserial correlations were calculated for: (1) a sample of postinstructed mastery students, and (2) a combined sample of postinstructed mastery and preinstructed nonmastery students. The mastery sample showed significantly less score variability than the combined sample. Haladyna's results indicated substantial correlations between the item indices calculated by D and the point biserial for the combined sample. The correlations between D and the point biserial for only the mastery sample were substantially lower. From this, Haladyna concluded that the point biserial coefficient provides an adequate item index when calculated on a combined sample of mastery and nonmastery students

From the above, it appears that the D measure and the point biserial coefficient for combined mastery and nonmastery groups are similar measurements. For the D index, an item is acceptable if preinstructed nonmasters fail the item while postinstructed masters pass the item. For the point biserial index, an item is acceptable if low total test scorers fail the item while high total test scorers pass the item.

DISCUSSION

The item analytic techniques reviewed suggest agreement on the general theme that a useful CR item is one which masters pass and which nonmasters fail. Table 2.1 presents the techniques reviewed with their respective authors and operational definitions.

The review failed to indicate any one item analytic technique which can be fully supported. Most measures assume, in one way or another, that a master is or should be the student who has been instructed. This will allow the item index to vary with instructional merit. It seems strange that no one, to our knowledge, has sought to remove this variance through variance analytic methods. Part correlation and partial correlation also come to mind in this regard.

There may also be a need for a unifying definition of a master. Such a definition, it seems, should be independent of the group to which it will be applied. Siegel, Schultz, Fischl, and Lanterman (1968) used a modified method of limits to derive such a definition. New and other creative approaches to this definitional problem seem required.

Table 2.1

Criterion-Referenced Item

Analytic Techniques

Part A: Item Validity Measures

Author	Design	Mastery Definition	Measure[ad]
1. Cox and Vargas (1966)	one group pretest and posttest	instruction[b]	$\left[\frac{a}{(a+b)} - \frac{c}{(c+d)}\right]$
2. Ivens (1970)	one group pretest and posttest	instruction	$1 - \left[\frac{(w+z)}{(w+x+y+z)}\right]$ or $\left[\frac{a}{(a+b)}\right] - \left[\frac{c}{(c+d)}\right]$
3. Rahmlow, Mathews, & Jung (1970)	two group instruction versus no instruction	instruction	$\left[\frac{a}{(a+b)}\right] - \left[\frac{c}{(c+d)}\right]$ together with $\left[\frac{a}{(a+b)}\right]$
4. Hambleton & Gorth (1971)	one group pretest and "delayed" posttest	instruction	$\left[\frac{a}{(a+b)}\right] - \left[\frac{c}{(c+d)}\right]$
5. Hsu (1971)	one group pretest or posttest	empirical cut-off score	$\left[\frac{a}{(a+b)}\right] - \left[\frac{c}{(c+d)}\right]$ or phi coefficient
6. Popham (1971)	one group pretest and posttest	instruction	rank correlation between $\left[\frac{y}{(w+x+y+z)}\right]$ and $\left[\frac{x}{(w+x+y+z)}\right]$
7. Crehan (1974)	two group instruction and no instruction	instruction	$\left[\frac{a}{(a+b)}\right] - \left[\frac{c}{(c+d)}\right]$

Table 2.1 (cont.)

8. Haladyna (1974)	two group instruction and no instruction	empirical cut-off score and instruction	point-biserial correlation for a combined mastery and nonmastery sample

Part B: Item Reliability Measures

Author	Design	Reliability Definition	Measure[d]
1. Ivens (1970)	one group posttest and a retest	stability	$\left[\frac{(w+z)}{(w+x+y+z)}\right]$ or absolute value of $\frac{(w+y)-(w+x)}{(w+x+y+z)}$
2. Popham (1971)	one group pretest and posttest	internal consistency	χ^2 goodness of fit test with the median frequency across items as the oretical values

Notes

a. The symbols are defined by the four cells of a mastery level (master or nonmaster) by item response (pass or fail) frequency matrix:

a = frequency within master by pass cell
b = frequency within master by fail cell
c = frequency within nonmaster by pass cell
d = frequency within nonmaster by fail cell

b. Pre or no instruction is defined as nonmastery; postinstruction is defined as mastery.

c. Instructed students who pass cut-off score are masters; noninstructed students who fail cut-off score are nonmasters.

d. The symbols are defined by the four cells in a frequency table of prepost item response patterns:

w = frequency of pass-pass responses
x = frequency of pass-fail responses
y = frequency of fail-pass responses
z = frequency of fail-fail responses

III. RELIABILITY AND VALIDITY

Test reliability is generally defined as the extent to which test scores are free from random error variance. Reliability, then, is seen as consistency of measurement. Traditionally, test reliability has been estimated through correlation statistics. Three established estimates of test reliability are: (1) coefficient of stability--correlation of measures with the same set of measures obtained at a later time, (2) coefficient of equivalence--correlation between measures obtained from equivalent instruments, and (3) coefficient of internal consistency--correlation obtained from an internal analysis of data collected on a single administration of the measurement instrument. Such correlational estimates of reliability, however, have been held to be inappropriate for CR tests. Given the reduced score variance, correlation coefficients become depressed so that a possible highly stable, equivalent, and internally consistent CR test could yield traditional reliability coefficients near zero. Accordingly, other reliability measures have been held to be needed and several have been developed to assess specifically the reliability of CR tests.

Although statistical procedures for use in estimating the reliability of CR tests are still mostly in the exploratory stages, approaches other than the traditional ones have been suggested: Ivens (1970), Popham (1971), Unks (1971), Livingston (1972), Hambleton and Novick (1973), Crehan (1974), Haladyna (1974), and Swaminathan, Hambleton, and Algina (1974).

Ivens' Stability and Equivalency Indices

Ivens (1970) presented two types of measure for assessing test reliability. One measure (f_1) was considered a stability index while the second (f_2) was considered an equivalency index. Each index was based on the proportion of people achieving the same score, or virtually the same score, on the postinstruction tests. For the stability index, the two tests are a test and a retest score using the same postinstruction test. For the equivalency index, the two measures are a test and a retest using two alternate forms of a postinstruction test. A score is considered the same from test to test if it varies less than a specified small value (say 5%). Each reliability index can range from zero to one, with one indicating perfect correspondence between scores. Ivens also suggested that the correspondence between scores be reported as the proportion achieving a retest score within a given percentage value of the earlier test. For example, reliability could be reported if 90% of the students' retest score were within 8% of the scores from the first test.

It is noted here that Ivens' (1970) item analytic measures (as well as those of the other authors previously discussed) might be considered to be an estimate of overall test reliability by determining an average (mean or median) across items. Ivens suggested such a step for his two item measures, f_1 and f_2.

Popham's Internal Consistency Measure

For an estimation of the internal consistency of a CR test, Popham (1971) suggested an approach which considers the overall similarity of item behavior. The approach is based on a pre and postinstruction testing paradigm. Each examinee's score on an item for both the pre and postinstruction test is categorized into one of the four possible response patterns: C-I, I-C, C-C or I-I where C = correct and I = incorrect. This procedure is performed for each item on the test. Hence, a 4 x N frequency table (where N is the number of items) is developed. A chi-square test of independence with 3(N-1) degrees of freedom is calculated. A statistically significant chi-square indicates that the items have different response patterns, and hence do not behave similarily.

Popham (1971) analyzed items from several subtests (each measuring a specific objective) with this chi-square approach. All but one chi-square value (from a total of 15) reached significance at the .05 level of confidence. Apparently surprised by such results, Popham indicated that the chi-square approach may lack utility. Popham's reasoning was that the items were constructed to measure the same objective, and accordingly, the chi-square values should not have reached significance.

Unks' Methods

Unks (1971) presented two novel methods for measuring CR test reliability. In both methods, a single group, pre and postinstruction test design is used. The first method calculates item-total score correlation coefficients for both the pretest and the posttest results. These correlations are the item validity coefficients traditionally used with NR tests. Unks then suggested that the correlation between the pretest and the posttest item validity coefficients provides a measure of test reliability. The approach measures the overall consistency of the relationship between item score and total test score. It apparently ignores the lack of within score variance characteristic of CR tests. If score variance is lacking, then the traditional item validity coefficients are restricted and less variable, thereby, reducing the correlation between the post and pretest item coefficients.

Unks' (1971) second approach to reliability measurement was based on the standard error of prediction; more specifically, on the standard deviation of the error of prediction which results when an item score is used to predict the total test score. A simple linear regression equation is used. Two sets of standard errors are calculated. One set is calculated for the item and total score from the pretest and the second set is caluclated from the posttest. The two sets of standard errors are then compared using a matched group t-test statistic. Unks suggested (it seems incorrectly) that the probability level of the resultant t-value serve as an index of reliability. The higher the probability level, the more reliable the test. This index can be taken to indicate the overall consistency of standard error in a test.

In evaluation, it appears that Unks' (1971) two reliability measures lack any logical relationship to the concept of test reliability as the concept is generally understood. As noted, test reliability is inversely related to the amount of random error in the measurement. And, an indication of random error is shown by the degree of consistency of the measurement. Anything random, by definition, does not occur consistently. With Unks, however, the crucial phrase is consistency of the measurements, (i.e., the instrument). Unks' measures apparently show a consistency, but not a consistency of the measurement instrument itself, rather a consistency of derived, reliability type measures. Thereby, the relationship of Unks' two test reliability measures to the extent of random error is indirect and vague.

Livingston's Traditional Measure Analogues

The problem posed by the lack of variability simply stated is that with zero variability many of the statistical terms employed in a correlational analysis become undefined. Livingston (1972) maintained that these problems are due to the use of NR definitions for CR tests. He redefined the concept of variance to increase its relevance and usefulness in the area of CR measures. The classic approach to variability views all variability as deviations from a central score (the mean). In CR testing the score of interest is not the mean but the criterion score. The criterion score is the pass-fail, master-nonmaster cut-off score in a CR test.

Livingston (1972), accordingly, redefined variance as the mean squared deviation from the criterion score. The equation reads as follows:

$$D(x) = \frac{\sum_{i=1}^{n}(X-Cx)^2}{n-1}$$

where D(x) is Livingston's analogue to the variance of X, Cx is the criterion score and X is the test score for the ith testee. (Livingston's notation is simplified here and below). Likewise, Livingston defined covariance in terms of deviations from the criterion score:

$$D'(X, Y) = \sum_{i=1}^{n} \frac{(X-Cx)\,(Y-Cy)}{n-1}$$

where D'(X,Y) is Livingston's analogue to the covariance. Accordingly, Livingston presented an analogue to the traditional product-moment correlation coefficient:

$$k(X, Y) = \frac{D'(X, Y)}{\sqrt{D(X)D(Y)}}$$

Given the definitions, Livingston (1972) developed a rationale for the measurement of CR test reliability. As with classical test theory, Livingston defined reliability as the squared correlation between the observed score and the true score. (Note the observed score equals the true score plus the error score). The equation reads:

$$k^2\,(X, Tx) = \frac{[D'(X, Tx)]^2}{D(X)\;D(Tx)}$$

where Tx is the true score for a person on X. Livingston showed that this equation equaled the ratio of the variance of true scores to the variance of the observed scores:

$$k^2(X, Tx) = \frac{D(Tx)}{D(X)}$$

Interpretation of the results of this equation is analogous to interpreting a reliability coefficient in classical test theory.

Livingston (1972) further showed that the analogy to classic test theory extends to the correlation between two parallel forms of a CR test which he showed to be equal to the CR reliability of either of the tests:

$$k(X, Y) = [k(X, Tx)]^2$$

In relating his CR statistics to those of classic NR testing, Livingston showed that the CR reliability correlation coefficient is a general case of the classic reliability coefficient, and that, indeed, the CR reliability coefficient is always at least as large as the classic formulation, specifically:

$$[k(X, Tx)]^2 \geq [r(X, Tx)]^2$$

where r(X,Tx) is the correlation between the observed and true score calculated using the classic NR approach. Livingston supplied an insight into the reason for this relationship using the following logic. The CR reliability coefficient reflects uncertainty relative to the criterion score, while the NR reliability coefficient reflects uncertainty with respect to the mean. In many cases, the scores will all be quite far from (all above or all below) the criterion score. In such cases there would be no uncertainty with respect to a score being above or below the criterion score, however, uncertainty with respect to the mean would remain. Only when the mean equals the criterion score is the uncertainty equal in the two approaches.

Livingston (1972) provided other important, but less definitional, CR analogs in NR test statistics, including: Spearman-Brown prophecy formula, coefficient alphs, and a correction for attentuation. In all, Livingston developed and presented statistics for calculating various types of test reliability for CR tests. These CR statistics may be used in calculating such common reliability measures as: test-retest reliability, parallel form reliability, and split-half reliability. Moreover, it would appear from Livingston's work that for item analytic purposes one could develop a CR analog to the point biserial correlation coefficient to index the relationship between an item and the total test score.

Livingston (1972) strongly suggested the use of his CR test statistics over the NR test statistics when measuring the reliability of a CR test. He attempted to show how the use of NR statistics for CR tests may be misleading. Specifically, it is possible for two tests to be shown as uncorrelated with NR statistics but as highly correlated with CR statistics. This may occur when all students: (1) score above the criterion on the two tests, and (2) the shape of the bivariate scatterplot of scores on the two tests in circular. Here, the NR correlation between the two tests is about zero; a student who scores high on one test, in relation to the mean, is not especially likely to score high on the other test, in relation to the mean. On the other hand, the CR correlation between the two tests is high and positive; a student who scores high in relation to the criterion on one test is likely to score high in relation to criterion on the second test.

Livingston (1972) apparently has provided an important contribution to the score variance problem in CR testing. By replacing the mean score with the criterion score, Livingston has developed test statistics specifically adapted to CR measurement. However, Livingston's approach has been criticized (e.g., Harris, 1972; Shavelson, Block, and Ravitch, 1972) by others. Shavelson, Block, and Ravitch (1972) presented evidence which showed the NR reliability coefficient and CR reliability coefficient as distinctly different measures. These authors suggested that Livingston's measure be given some other name than "reliability."

Hambleton and Novick's Replicability Concept

Hambleton and Novick (1973) noted that the concept of reliability is in essence replicability. That is, the less the random error, the more consistent and more repeatable the measurement result and hence the greater the reliability. Thereby, reliability measures need not depend on score variance as replicability is not, in essense, dependent upon score variance. To Hambleton and Novick, the reliability of a test (either a CR or NR test) can be estimated by a nominal scale comparison of test results (e.g., pass-fail, master-nonmaster). Specifically, Hambleton and Novick suggested that test reliability (replicability) could be estimated by comparing the results of a test administered to two comparable groups. Or alternatively, reliability could be estimated by comparing the results of parallel tests administered to the same group. The percentage of passing scores on each of the test administrations can be used to compare test results. The more comparable the percentages, the more reliable the test.

Haladyna's Internal Consistency Measure

Haladyna (1974) maintained that if sufficient variance exists when mastery and nonmastery CR scores are combined, then internal consistency coefficients may be properly estimated. He stated that the degree of homogeneity for any CR test, given combined samples, appears to be an empirical issue. In an experiment, Haladyna found that the variance was significantly greater when he combined pre and postinstruction group test scores than the postinstruction group's variance by itself. He examined internal consistency with the KR-20 formula and found that the combined group sample yielded higher estimates of homogeneity (internal consistency) than the uncombined estimates. Increasing the variability by combining mastery and nonmastery student's test results yielded correspondingly higher reliability estimates. Based on these findings, Haladyna concluded that internal consistency is a workable concept for determining the reliability of CR tests, given that pre and postinstruction test scores are employed.

Swaminathan, Hambleton, and Algina's Decision Consistency Approach

Swaminathan, Hambleton, and Algina (1974) presented a decision-theoretic approach to the measurement of CR test reliability. They considered the consistency of decisions about mastery states (e.g., master or nonmaster) on repeated administrations of the test as a measure of test reliabiliy. Specifically, they defined reliability of a CR test as the measure of agreement between the decisions made (i.e., the student passed or failed the test) in repeated test administrations.

The measure of reliability is given by the coefficient of agreement K:

$$K = [(Po-Pc)/(1-Pc)]$$

where Po is the observed proportion of agreement between decisions and Pc is the expected proportion of agreement between decisions. Po is given by:

$$Po = \sum_{i=i} Pii$$

and Pc is given by:

$$Pc = \sum_{i=1} Pi.\ P.i$$

where Pii equals the proportion of examinees placed in the ith mastery state on both test administrations, and Pi. and P.i represent the proportion of examinees assigned to the mastery state i on the first and second test administrations, respectively.

To illustrate this approach, suppose that examinees were classified into two mastery states, master or nonmastery, based partly on the test results of each of two test administrations. These data are then cross classified and joint, marginal, and expected proportions calculated--much as performed for frequency data in the chi-square goodness of fit test. Some hypothetical proportions are presented in Table 3.1.

Using the proportions presented in Table 3.1:

$$Po = .70 + .19$$
$$= .89$$

$$Pc = (.75)(.76) + (.25)(.24)$$
$$= .63$$

and

$$K = (.89 - .63)/(1 - .63)$$
$$= .70$$

In the above example, the value of K is .70, which suggest a fair amount of consistency--as would be anticipated from the data. In theory, the upper limit of K is +1 and may occur when the marginal proportions for different test administrations are equal. A K value less than 1 will occur if any examinee is classified differently on repeated administrations. The lower limit of K approaches -1. Any negative K value shows extreme decision making inconsistency and unreliability. However, a linear interpretation of specific negative K values may be misleading. As inconsistency increases across classification decisions, K values may not decrease correspondingly. Consider the examples shown in Table 3.2. Table 3.2a presents proportions inverted to those shown in Table 3.1. A K = - .41 results:

Table 3.1

Summary Information for Hypothetical Data on the Joint Classification of Examinees into Two Mastery States on Two Test Administrations

Admin 1 \ Admin 2	Mastery States: Master	Mastery States: Non Master	Marginal Proportion
Mastery States: Master	.70(.57)*	.05	.76
Mastery States: Non Master	.05	.19(.06)*	.24
Marginal Proportion	.75	.25	

*Expected proportion

Table 3.2

Summary Information for Hypothetical Data for Various Examples of Internal Consistency Measure

a.

Admin 1	Admin 2 Mastery States: Master	Non-Master	Marginal Proportion
Mastery States: Master	.05(18)*	.19	.24
Mastery States: Non-Master	.70	.06(.19)*	.76
Marginal Proportion	.75	.25	

K=-.41

b.

Admin 1	Admin 2 Mastery States: Master	Non-Master	Marginal Proportion
Mastery States: Master	.95(.90)*	.00	.95
Mastery States: Non-Master	.00	.05(.00)*	.05
Marginal Proportion	.95	.05	

K=1.00

c.

Admin 1	Admin 2 Mastery States: Master	Non-Master	Marginal Proportion
Mastery States: Master	.00(.05)*	.05	.05
Mastery States: Non-Master	.95	.00(.05)*	.95
Marginal Proportion	.95	.05	

K=-.11

*Expected proportion

this is not symmetrical to that produced from Table 3.1. Tables 3.2b and 3.2c present more extreme proportions inverse to one another. These show perfect classification consistency and inconsistency respectively, across the two test administrations. With the perfect consistency shown in Table 3.2b, K equals a perfect +1 (as it should). On the other hand, with the perfect inconsistency shown in Table 3.2c, K equals the quite less than perfect -.11. This last result is neither symmetrical nor linear; if it was symmetrical one would expect a -1. Accordingly, K, unlike a typical reliability coefficient, may not possess interpretive clarity. This result suggests the need for investigating the operating characteristics of novel statistics studied in this area prior to their application.

The coefficient of agreement, K, as a measure of CR test reliability, may possibly be taken to indicate the proportion of agreement between decisions that exist, over and above that which can be expected by chance alone. In this regard, Swaminathan et al., maintained that a measure which shows the percentage of examinees placed in the same mastery state over two test administrations is lacking because such a measure does not take into account the fact that agreement could occur by chance alone. Such chance agreement confounds a reliability measure.

We also note that the K coefficient is not solely concerned with the reliability of a CR test, itself. The K coefficient, as presented, is dependent on the entire decision-making process. Such a reliability measure then, is concerned not only with the content of a CR test but also the decisions made on the basis of the test scores.

Summary

Several approaches to measuring the reliability of CR tests are available. All suffer from one or more conceptual or statistical drawbacks. There appears to be no agreement on a preferred approach. Part of the problem may lie in the desire to mimic NR tests when CR tests are under consideration. Another issue seems to be the type of reliability that is important for CR tests. Why should CR reliability march to the music of NR reliability? Perhaps CR reliability hears a different drummer. Table 3.3 presents an outline of the CR test reliability measures reviewed.

Table 3.3

Summary of CR Reliability Measures

Author	Design	Reliability Definition	Measure
1. Ivens (1970)	one group test-retest	stability or equivalency	proportion of test scores that remain the same or virtually the same
2. Popham (1971)	one group pre- and post-instruction test	internal consistency	χ^2 test of independence with item response patterns as one demension and the items themselves as the second dimension
3. Unks (1971)	one group pre- and post-instruction test		correlation between the item validity coefficients on the respective tests or correlation between the standard errors for the items on the respective tests
4. Livingston (1972)		stability equivalency or internal consistency	uses correlations which are analogs to classic test reliability statistics: the criterion score replaces the mean
5. Hambleton and Novick (1973)	one group, test-retest	stability or equivalency	proportion of examinees passing each of the tests
6. Haladyna (1974)	two group instruction and no instruction	internal consistency	KR-20 coefficient computed on combined mastery and nonmastery group
7. Swaminathan, Hambleton, and Algina (1974)	one group test-retest	stability or equivalency	proportion of examinees where the same mastery decision is made beyond chance

CR TEST VALIDITY MEASURES

Guion (1965) stated that while test reliability may be the sine qua non of testing (i.e., if a test is unreliable, it can not have any merit), acceptable reliability alone in insufficient. Evidence of test validity is a necessary requisite for establishing whether or not a test is a "good" measure.

The Standards for Educational and Psychological Tests (American Psychological Association, 1974) defined validity as the appropriateness of the inferences made from test scores or other forms of assessment. In all, four aspects of validity are generally considered when NR tests are involved: (1) predictive, (2) concurrent, (3) construct, and (4) content validity. Predictive, concurrent, and, to some extent construct validity are empirical types of validity. That is, quantitative data are acquired and analyzed in order to assess validity. Content validity is a more logical, judgmental type of validity. A test is content valid to the extent its items are judged to represent the domain of the testing objectives. For CR tests, it is with the empirical validities that problems arise.

Since the procedures typically employed for empirically assessing test validity are correlational in nature, they are based on score variance. Popham and Husek (1969) noted relative to CR tests "...the results of the procedures (empirical validation) are useful if they (correlations) are positive, but not necessarily devastating if they (correlations) are negative." (p.6) Predictive validity measures, in theory, provide an indication of the effectiveness of a test for predicting an individual's behavior in specific situations. For this purpose, performance on a test is usually checked against an external criterion, an independent measure that the test is designed to predict. Predictive validity, however, for the CR test has been considered irrelevant (e.g., Gagne, 1969; Ivens, 1970; and Shriver and Foley, 1974). Instead, content validity, with its concern for what is being measured, is the validation strategy most often used by CR test developers (Popham and Husek, 1969; Ivens, 1970; Gagne, 1969; Klein and Kosecoff, 1973; Gronlund, 1973; Sweezey and Pearlstein, 1975).

Content validity for a CR test may be determined by: (1) systematically developing the test. (i.e., referencing the test items directly to criterion objectives), (2) obtaining expert judgment of the appropriateness of each item for measuring mastery of an objective, and (3) item analyzing the test to determine if the test items correlate more highly with other items used for measuring the same objective than they do with items used for other objectives (Klein & Kosecoff, 1973),* If the test items match the objectives precisely, the test is content valid. The measurement is observational and rests entirely on the judgment of experts (Sweezey and Pearlstein, 1975).

---

*Klein and Kosecoff (1973) noted that item analysis will suffer from lack of test variance.

Although many CR test developers support the value of content validity for determining the validity of the CR test, and in the opinion of many this type of validity is the only one that should be considered, more empirical validity measures have been considered. These validity approaches, as well as the content validation strategy, are reviewed below. The authors included are: Popham and Husek (1969), Ivens (1970), Cox (1971), Unks (1971), Shriver and Foley (1974), and Sweezey and Pearlstein (1975).

Popham and Husek's Construct Approach

As noted, Popham and Husek (1969) suggested that the validity of a CR test may be indexed using classic measures but that the results are useful only if they are positive, and are not clearly interpretable if negative. They maintained that CR measures are primarily validated in terms of content validity-the extent to which the test resembles the criterion. However, construct validity approaches are also seen as appropriate for CR tests. Construct validity is useful in the case of a test which measures "a proximate predictor (e.g., administred at the close of instruction) of some more distal criterion (e.g., occuring many years hence)". A positive intercorrelation among several such scores indicates the presence of construct validity according to Popham and Husek. That is, empirical evidence for the construct validity of a CR test may be inferred from the intercorrelations of several proximate predictors of the same ultimate criterion.

Ivens' Gain Score

Ivens (1970) suggested that the validity of a CR test may be determined by the "magnitude of the gains shown between the pretest and posttest means...In the comparison of two tests, it is plausible to assume that the one with the largest gain from pretest to posttest is the most adequately reflecting subject proficiency on the stated objective.", (p. 13). Standard statistical tests for determining the significance of differences may be applied to the results of the calculations to determine if the gains are significantly different from chance increments. Such a statistic might show evidence for or against construct validity.

As discussed previously, Ivens suggested two measures for evaluating CR items. To repeat, the quantitative technique requires three administrations of the same test to the same examinees--a pretest(a), a posttest(b), and retest(c). Recall the first index is:

$$f_1 = (1 - P_{ab})(P_{bc})$$

P_{ab} and P_{bc} are defined as the proportion of examinees whose item scores are identical over the two subscripted administrations. The maximum score of 1 is obtained if all the examinees fail the item on the pretest and pass the item on the posttest and retest (in the pass/fail scoring format).

The formula for the second index is:

$$f_2 = (P_b - P_a)\,(1 - |P_c - P_b|)$$

where P is the proportion of examinees who pass a given item. In the second formula the range varies from -1 to +1. Negative values occur if a greater proportion of examinees pass the item on the pretest(a) than on the posttest(b). These formulas, which assess item quality, can be used to assess overall test validity by calculating mean values over all items. As noted, these indices may be used as substitute measures for test reliability. The first term in both the formulas was held to be a measure of item validity and when averaged across items to provide a type of internal consistency estimate. For test validity purposes, evidence of construct validity is shown by the degree of internal consistency. The second term in the formulas are measures of item reliability and show stability when averaged across items. And finally, Ivens suggested that predictive validity can be measured by substituting independent measures of actual objectives for the retest.

To the extent that these arguments hold, the various item indices and the internal consistency test reliability measures discussed in the previous sections can be used for evidence of construct validity.

Cox's Construct Approach

In an approach similar to that suggested by Popham and Husek (1969) and actually employed by Ivens (1970), Cox (1971) stated that a construct validation approach may be successfully applied to CR tests. As an example of this, Cox suggested the use of a comparison of pre and postinstruction test scores as a measure of CR test validity. Such a measure would provide an estimate of internal consistency, and hence can be taken as evidence for construct validity. However, Cox did not provide the details for the application of this approach and no specific measure was given.

Unks' Concepts

Unks (1971) identified three viewpoints on measuring the validity of CR tests. One was Ebel's (1961) view that the CR test sometimes can not be further validated, since the test itself is the best available definition of the criterion. Ebel's view indicates that CR tests are content valid by definition. A second approach involved the creation of sequentially scaled items to form a test with Guttman scale properties. An evaluation of the extent to which the actual data show this

property is taken as an indication of test validity. A third viewpoint, presented by Unks, considered the standard deviation of the error distribution (i.e., the error in predicting the criterion from the test scores: the standard error) as a validity estimate. This final approach may use simple linear regression with an external and independently measured criterion as the dependent variable and the CR test scores as the independent variable. Such an approach represents a predictive validation strategy.

Shriver and Foley's Job Relevance

Shriver and Foley (1974) developed a battery of CR performance tests for several types of job activities required for electronic maintenance (e.g., align, adjust, and calibrate; remove and replace; and use of hand tools). This approach emphasized test construction and the development of test tasks which directly referenced the job activities. As such, content validity was established. Indeed, Shriver and Foley (1974) stated that because they designed their CR test to be as nearly identical to job criteria as possible, "...no validation of their empirical validity is possible. They are empirically valid by definition, (p. 44)." Furthermore, Shriver and Foley noted that their performance type CR tests measure only skills and abilities and not motivational variables. Thereby, test performance will predict job performance only within limits and this predictive validity is reduced.

Shriver and Foley's contention seems defensible to the extent that: (1) the test and the ultimate job are congruent, and (2) the job analysis is fully descriptive and generalizable.

Sweezey and Pearlstein's Concurrent Validity

Sweezey and Pearlstein (1975) discussed concurrent validity for CR tests. To obtain concurrent validity, the CR test results of individuals are compared with their results on other measures of performance (taken in close proximity to the test). For CR tests, the other measure would, of necessity, have to be an independent assessment of performance on the same criterion objectives. A phi correlational analysis (which is less dependent a score variance than other correlational statistics) of these data was suggested to provide an indication of the degree of relationship, or concurrent validity. And, a t-statistic may be used to compare the mean score on the independent performance measure of the high (pass) and low (fail) test scoring groups to show whether or not a concurrent relationship exists.

In addition, Sweezey and Pearlstein (1975) discussed predictive validity. According to Sweezey and Pearlstein, predictive validity for a CR test rests on the same concept as the concurrent validity. Predictive validity compares CR test results with the results of another measure, taken later in time, usually a measure obtained when the students are on-the-job. Criterion measures for predictive validity purposes are: supervisor ratings, other tests, peer ratings, other on-the-job measures of performance (e.g., time on-the-job until proficiency was reached, level of productivity, errors made, number of times supervision was required). The phi correlation coefficient was also suggested by Sweezey and Pearlstein for obtaining a predictive validity index.

Summary

The present discussion considered four types of validity related to CR tests: (1) content, (2) construct, (3) predictive, and (4) concurrent. The various approaches are summarized in Table 3.4.

Several authors maintained that content validity is the only relevant aspect of validity for CR tests. This viewpoint was held not only because the lack of score variance inhibits the other correlational aspects of validity but also because content validity is primarily concerned with what CR testing is all about: the content meaning of a test. These arguments seem strong. However, the emphasis on content validity to the exclusion of other validation approaches may be considered to represent ignoring the problem. Validity must be both empirical and judgmental. As such, both empirical and jugmental answers are needed and necessary.

The lack of score variance in CR tests makes the validational problem difficult as was also true for the item analytic problem. The respective reviews of item analysis, test reliability, and validity presented several methods for managing the score variance problem. Three general procedures can be abstracted: (1) proportional and frequency type analyses (e.g., Cox and Vargas, 1966; Ivens, 1970), (2) CR analogs to the traditional correlational analyses (e.g., Livingston, 1972), and (3) traditional correlational analyses (e.g., Haladyna, 1974). The proportional and frequency-type analyses, as a whole, avoid the score variance problem and typically fit the CR testing paradigm.

Generally, there is confusion relative to the "proper" statistic in each area of concern. Most of the individual studies have been fragmentary and isolated. There seems to have been more interest in developing CR tests than in empirically evaluating them. User demand may constitute one reason for this. The current EEOC emphasis on job relevance gives additional thrust to the content arguments. And, as EEOC requirements move away from validational concepts towards nonadverse impact requirements, the formal statistical aspects of a test, whether CR or NR, may receive further deemphasis.

Table 3.4

Summary of CR Validity Measures

Author	Validity Approach	Method	Criterion	Measure
1. Popham and Husek (1969)	construct	empirical	external	correlational
	content	logical		judgmental
2. Ivens (1970)	construct	empirical	internal	difference score
	predictive	empirical	external	difference score
3. Cox (1971)	construct	empirical	internal	
4. Unks (1971)	content	logical	-	
	-	empirical	-	Guttman scale
	predictive	empirical	external	regression: standard error of measurement
5. Shriver and Foley (1974)	content	logical	-	judgmental
6. Sweezey and Pearlstein (1975)	concurrent or predictive	empirical	external	correlational: phi coefficient

IV. MASTERY DETERMINATION/DEFINITION AND TEST LENGTH

Determining mastery involves two problems. First, it is not practical to insist on perfect test scores. Accordingly, complete mastery is almost nonexistent. Second, including in a test all items from the population of test items is often impossible. For example, in an addition test, from the indefinitely large population of addition items, a sample must be selected. Given these two factors, mastery determination must be made without perfect knowledge. That is, determining mastery requires a cut-off score at which a certain minimal number or percentage of the sampled items (selected for the test) are passed.

Parenthetically, we note that some CR testing paradigms include all the possible items from the item population and insist on perfect mastery. Such a paradigm more often than not involves performance testing. For example, Shriver and Foley (1974) developed a CR performance test battery which mirrored the criterion objectives. Their items included all the behaviors performed on the job. Second, Shriver and Foley insisted on complete mastery. In such a case no mastery extrapolation is required; the examinee can or can not successfully perform the whole job.

On the general level, Millman (1973) presented four considerations relative to determining cut-off scores:

- Item Content--The item content concept involved subjective judgments of how important it is that each item be answered correctly. On the basis of these evaluations, a minimum score for passing the test is determined. Or, the decision could be reached that, to pass the test, all items must be completed correctly. Alternatively, this rubric would have the test items classified in a matrix, with difficulty and importance as the dimensions. Here, Millman indicated that judgments are made of the proportion of items in each cell of the matrix that must be passed to be minimally qualified. The sum across the cells according to Millman, is the number of items which msut be answered correctly to pass the test.

- Educational Consequences--Millman said that the educational consequences concept considers the effects on future learning as a mastery determinant. If the mastery level is set too low, students may be given instruction on new concepts and skills they will be ill equipped to master. If the level is set too high, efficiency will be reduced since students will be spending too much time in remedial training.

- Psychological and Financial Costs--Millman described the psychological and financial cost concept as dictating a low cut-off score if the psychological and/or financial costs are high. Examples of psychological costs are reduced motivation, boredom, and damage to one's self-image.

- Errors of Guessing and Item Sampling--Millman's guessing error and item sampling error concept considers score adjustment to account for guessing or item sampling errors. Millman recommended that the cut-off scores be raised or lowered when the test items do not fully represent the population of behaviors covered in the unit of instruction, so that misclassifying students does not occur.

Deciding whether or not a student has passed a test and is a master of some performance has been accomplished within varying degrees of arbitrariness. At a more arbitrary level, a cut-off score has been selected on the basis of "best judgment" of knowledgeable people--people who represent the training and job requirements points-of-view. At a less arbitrary level, cut-off scores have been selected on the basis of a statistical model which attempts to minimuze errors in mastery classifications, (i.e., classifying a "true" master as a nonmaster or a "true" nonmaster as a master). The present chapter reviews certain of these methods for selecting cut-off scores.

Delphi Technique

Siegel, Bergman, and Lambert (1973) employed the Delphi technique to set minimally acceptable and desirable test scores. The Delphi technique, developed at the Rand Corporation, (Dalkey and Helmer, 1962; Helmer, 1967; Dalkey, 1967; Brown, 1968; Dalkey, 1969; Martino, 1972) is a method for converging the opinions of a group. Siegel, Bergman, and Lambert (1973) asked supervisors to provide a quantitative estimate of the cut-off score on a CR test which would define a mastery demarcation. The judgments made by the experts were individually determined, without benefit of consultation with the other group members. The estimates were then collected and presented to the group, as a whole, so that each could see his estimate in the context of the other group member's estimates. Then, a session was conducted in which various supervisors were asked to justify their estimate. Following the justification procedure, the supervisors reassigned cut scores. This procedure was followed until the group reached a consensus.

Statistical Models

Several more statistical approaches to the selection of cut-off scores have also been developed. Epstein, Steinheiser, Macready, and Mirabella (1977) provided an extensive review on the statistical models for determining mastery level. Some of the models which in our view are most practical, are described below. The models are categorized for convenience as: (1) empirical, (2) probabilistic, (3) binomial and (4) Bayesian.

Empirical Models

Block (1972) and Crehan (1974) developed empirical methods for establishing cut-off scores.

Block (1972) established cut-off scores on the basis of the relationship of test scores to a set of external criteria. His criteria considered performance and attitudinal variables. Specifically, Block experimentally examined the effect of using different cut-off scores--0% (the control group), 65%, 75%, 85%, or 95% correct--on five separate outcome measures: (1) achievement, (2) learning rate, (3) transfer, (4) interest, and (5) attitude. From Block's work, it appears that the test score which optimally discriminates between high and low performers on the external criteria can be selected as the cut-off score. This approach takes into account the attitudinal as well as the performance criteria; that is, a tradeoff is made by selecting a cut-off score which optimally discriminates on both the performance variables and the atitudinal variables. The major problem with Block's approach seems to be that the results may not be generalizable from sample to sample. Continuous cross validation may be necessary.

Crehan (1974) presented a method for setting cut-off scores based on instructional effects. This method is based on the relative success of possible cut-off scores in discriminating between pre and postinstruction students. Accurate discrimination is assumed when preinstruction students fail and postinstruction students pass the tests. Specifically, Crehan's approach chooses that test score which maximizes the number of: (1) preinstruction students who fail the test, and (2) postinstruction students who pass the test. As noted in a previous section, Crehan defined mastery as dependent on instruciton. And, also as noted previously, such a method is confounded to the extent that: (1) mastery is achieved independent of formal instruction, and (2) instructional quality varies.

Probabilistic Approaches

Emrick (1971) and Dayton and Macready (1976) presented two related probabilistic models which estimate the optimal mastery level

cut-off score. Both models assume that mastery level is all-or-none; one is either a master or a nonmaster with respect to some skill. Essentially, these models choose that test score which minimizes the cost and probability of misclassification. The probability of misclassification is defined as the probability that a master will not achieve the cut-off score plus the probability that a nonmaster will equal or exceed the cut-off score.

Emrick (1971) presented a probability formulation for identifying an optimal cut-off in terms of a cost tradeoff between classifying a "true" master as a nonmaster (i.e., false negative) and a classifying a "true" nonmaster as a master (i.e., false positive). The optimized cut-off score formula is:

$$k = \frac{\left[\log\left(\frac{b}{1-a}\right)\right] + \left[\frac{1}{n}\log\frac{L_2 p(M)}{L_1 p(\bar{M})}\right]}{\left[\log\frac{ab}{(1-a)(1-b)}\right]}$$

where:

- k = cut-off score; percentage of items correct
- a = probability of "guessing" correct answer
- b = probability of "forgetting" correct answer
- $p(M)$ = probability of mastery
- $p(\bar{M})$ = probability of nonmastery
- L_1 = loss incurred from false positive
- L_2 = loss incurred from false negative
- n = test length, number of items

Dayton and Macready (1976) developed a probabilistic model for determining cut-off scores which accounted for "guessing" and "forgetting" errors. These errors are a result of the all-or-none, dichotomous mastery level assumption. An error occurs when a "true" master forgets the correct response or a "true" nonmaster guesses the correct response. In essence, the Dayton and Macready model estimates the probability of all the possible response patterns for a test given the: (1) probability of mastery, (2) probability of "guessing" correctly, and (3) the probability of "forgetting." That test response pattern which minimizes the probability of misclassification is chosen as the cut-off score. Specifically:

$$p(j) = [a^{sj}(1-a)^{n-sj}\, p(\bar{M})] + [b^{n-sj}(1-b)^{sj} p(M)]$$

where:

- $p(j)$ = probability of given response pattern
- sj = number of correct responses (e.g., number of 1's in the response pattern
- a = probability of "guessing"
- b = probability of "forgetting"
- $p(M)$ = probability of mastery
- $p(\bar{M})$ = probability of nonmastery
- n = number of items

We note that both Emrick (1971) and Dayton and Macready (1976) assumed that: $[p(M) + p(\bar{M}) = 1]$. That is, if mastery level is not all-or-none and partial or overlearning can occur, then the cut-off score determined by the two formulas will be less than optimal.

Either of the two approaches can also be used to determine the number of items required for a test (given the other variables in either equation).

Shoemaker (1972) suggested that a multiple cut-off method be employed with CR tests. With this approach, items are selected so that: (1) a certain proportion of items would be passed by all examinees reaching a minimum level of satisfactory achievement, (2) an additional proportion of items would be passed by those examinees who have surpassed the minimum level of achievement, and (3) the remaining test items would be passed by those examinees achieving a high level of mastery on that objective. It is possible to bracket a student's achievement level on each objective, with such a distribution, and still have a CR test. The three levels of mastery identified in this paradigm are the minimum levels, above minimum, and high level of achievement.

Binomial Model

Millman (1973) presented a binomial distribution model from which a cut-off score can be chosen such that the probability of masters not scoring at least that score and nonmasters scoring at that score is minimized. In this model, mastery level is defined as the probability that a person will respond correctly to a randomly selected test item from a specified population of items. Hence, as probability ranges from zero to one, the model holds a continuous, incremental view of learning. The model is represented by a binomial formula:

$$f(x) = \binom{n}{x} p^{x} (1-p)^{n-x}$$

where:

x = the total number of correct responses

$f(x)$ = the probability of test score x

n = the number of items

$\binom{n}{x}$ = the binomial coefficient: $\frac{n!}{x!\,(n-x)!}$

Millman's model can also be used to estimate test length given: (1) an acceptable level of classification error, (i.e., false positives and false negatives), and (2) an operational definition of masters and nonmasters. To determine test length, two values of x are specified: the lowest percentage score which will be accepted as defining mastery (Xm) and the highest percentage score which will be accepted as defining nonmastery (Xnm). (A value between these two scores is taken to indicate ambiguity with regard to mastery level). Next, the probability of: (1) misclassifying a master as a nonmaster (Mc), and (2) misclassifying a nonmaster as a master (NMc) is specified. Finally, a percentage cut-off score (c) is selected. The aim is to determine the test length (n) such that Mc and NMc are not exceeded given Xm and Xnm.

To illustrate, it is desired that the cumulative binomial probability of master scoring below a percentage cut-off score (c) be less than Mc:

$$P(x \leq c-1) \leq Mc$$

and hence,

$$\sum_{x=0}^{c-1} \binom{x}{n} p^x (1-p)^{n-x} \leq Mc$$

where p = Xm, the minimal mastery level. Similarly, it is desired that the probability of a nonmaster scoring above a cut-off score (c) be less than NMc:

$$P(x \geq c) \leq NMc$$

and hence,

$$\sum_{c}^{n} \binom{n}{x} p^x (1-p)^{n-x} \leq NMc$$

where p = Xnm, the maximal nonmastery level. Each equation is then solved for n. The minimal n which solves both equations is taken as the preferred test length. Millman provided tables which aid in solving the equations for n.

The major weakness of the binomial model is the requirement that mastery and nonmastery be operationally defined. The solutions provided by the model are incorrect to the extent the operational definition of the mastery level is incorrect. In such a case, Epstein and his collegues considered the model's solutions as conservative. More important is the point that, if the true mastery levels are known, there is no need for models to determine cut-off scores. The model, therefore, is somewhat circular in determining mastery level. Epstein et al, therefore suggested that the binomial model is most useful for approximations of test length and a cut score prior to collecting test data.

Bayesian Model

Epstein and Steinheiser (1975) used Bayes' Theorem for determining mastery level. Essentially, this model asks: Given past test performance, what is the probability of mastery or nonmastery? It is a conditional probability statement based on the condition of prior information. As for the binomial model, the Bayesian model does not assume learning to be all-or-none. Mastery level is defined as the probability of responding correctly to any random item from the item universe.

The Bayesian model can be used for deriving test length and cut-off score, if information about the quality of the examinee population is known prior to testing. Two pieces of prior information are required: (1) an estimate of the mastery levels possessed within the examinee population, and (2) the conditional probability that a randomly sampled item would be correctly answered given a specific mastery state.

The Epstein and Steinheiser (1975) model is based on a two-step algorithm. The first step yields the probability of an examinee being in a mastery state, i, given an item score t. The equation is:

$$p(M_i | t) = \frac{p(t | M_i)\, p(M_i)}{\sum_{i=1}^{s} p(t | M_i)\, p(M_i)}$$

where:

s = the number of mastery states,

t = the item score, (0 or 1)

Mi = the mastery state being considered

p(Mi) = the prior probability that a person is in mastery state i

p(t|Mi) = the probability of the score t, given the mastery state

The second step combines the p(Mi| t) probabilities for each item and yields a final probability of being in mastery state i given the total test score T. The equation is:

$$p(M_i | T) = \frac{\prod_{j=1}^{n} p(M_i | t_j)}{p(M_i)^{n-1}} \Bigg/ \frac{\sum_{i=1}^{s} \prod_{j=1}^{n} p(M_i | t_j)}{p(M_i)^{n-1}}$$

where:

j = 1,2,...n = the number of items

T = the total test score

In all, the Bayesian model can enhance the determination of mastery levels--given the availability of accurate prior information. Specifically, it can be shown that if accurate prior probabilities are available, then the Bayesian model can be used to achieve a given level of mastery level classification accuracy with fewer items than otehrwise possible. However, the prior probabilities concerning the mastery level of examinees can often be inaccurate. Typically, the prior probabilities are estimated on the beliefs and expectations of the examiners about the examinee population and thereby, subject to error.

Discussion

The literature indicates that mastery can be defined arbitrarily or systematically: Table 4.1 summarized the basic approaches used to set mastery cut-off scores. The choice of one approach over another will probably depend on the specific user environment. The decision process, whether clinically or statistically based, should consider the necessary variables and the consequences of an improper consideration. One consideration is test length.

Test Length

The final topic discussed in this chapter concerns test length. In determining mastery level, the question of test length is a crucial one. Indeed, one method for enhancing the determination of mastery or nonmastery and reducing the chance of misclassification is to increase the number of test items. However, increasing test length is not always practical and some statistical models (e.g., binomial) estimate the smallest number of items needed to determine mastery level.

Traditionally, test reliability is a function of test length. For practical purposes, the CR test should be as short as possible and yet test a criterion objective sufficiently well that judgments of mastery may be made. Although novel to the area of CR testing, sequential analysis--a method of testing over two decades old (e.g., Anastasi, 1953; Tiffin & Hudson, 1956; Wald, 1947)--could be brought to the scene in the interest of saving test time. Sequential testing would allow the CR test to be shortened for the proficient or the poor student.

Sequential testing is based on the premise that very high and very low proficiency can be detected with less than the full complement of test items. More extensive testing is necessary for students of borderline proficiency. Typically, a test item is administered based on the student's response, one of three decisions is

Table 4.1

Summary of Approaches for Setting Cut-off Scores

Approach	Nature of Learning	Measure	Outcomes Considered	Cut-off Score Specification
Expert judgment	undefined	subjective opinion(s)	not specified	cut-off score judged as the best
Delphi technique	undefined	empirical rating(s)	not specified	cut-off score on basis of group consensus
Empirical: Block (1972)	undefined	empirical relationship with external criteria	performance and attitudes	cut-off score that optimally discriminates among examinees on both performance and attitudinal criteria
Empirical: Crehan (1974)	undefined	empirical relationship with internal criteria		cut-off score that maximally discriminates between pre and post instructed examinees
Probabilistic: Emrich (1971) Dayton & Macready (1976)	all-or-none	probability of mastery and nonmastery together with probability of guessing and forgetting errors	test length costs of misclassification	cut-off score that minimizes the costs of misclassificatior
Multiple cut-off scores	incremental	empirical relationship with criteria	achievement	multiple cut-off scores which separates examinees into minimum, above minimum and high level of achievement
Binomial: Millman (1972)	continuous	probability of a test score given the probability of mastery and nonmastery	test length classification error	cut-off score that minimizes the chance of masters not scoring and nonmasters scoring at that score
Bayesian: Epstein & Steinheiser (1975)	continuous	probability of being in a mastery state given the final test score	test length	cut-off score that meets the posterior probability given the mastery state

reached: (1) continue testing (decision indeterminate), (2) accept (master), or (3) reject (nonmaster). If the first decision (continue testing) is reached, the testing process is continued until a proficiency classifiction can be made, or the test is terminated. Knerr and Epstein (1976) indicated that sequential testing requires approximately half as many items as tests of fixed length to achieve the same level of accuracy.

Along a philosophy similar to sequential analysis, CR tests may also be tailored to fit better the ability of each examinee. Tailored testing includes a number of different testing strategies for adapting the difficulty of test items to the testee's ability level, (Lord, 1970); it involves administering test items of a difficulty level which more closely matches each examinee's ability to perform the task under test. As with sequential testing, a tailored test may reduce the total number of test items administered and possibly test time without loss of reliability.

One type of tailored testing strategy is the stratified adaptive computerized testing model (stradaptive). In stradaptive testing, a computer is programmed to select items from different strata of item difficulty levels based on: (1) an initial estimate of the examinee's ability, and (2) the examinee's response to the previous item. The selected item is presented to the examinee (via a cathod-ray-tube) and his response recorded. At the start of testing, the difficulty level of the first item presented is based on the prior estimate of the examinee's ability. The next item presented will be: (a) more difficult if the response was correct, (b) less difficult if the response was incorrect, or (c) equal in difficulty if no response was made. Typically, this process continues until the examinee responds incorrectly or not at all to some defined number of items (e.g., 4 or 5) within a given stratum of item difficulty.

Waters (1977) presented evidence which suggested that a stradaptive version of a standardized ability test is more reliable, equally valid, and requires fewer items. Waters noted, however, that stradaptive testing required a longer response time per item, and suggested that stradaptive testing may not always be more efficient than conventional testing.

Millman (1973) added two concepts to aid in determining test length. One method uses the standard error of measurement which ultimately converts numerical values of standard errors into probability statements. The other method was referred to as the bimodal model. Tables were presented which related test length to accuracy for a given passing score out of a total number of items for student's true level of functioning.

Sweezey and Pearlstein (1975) suggested several general approaches for determining the number of items a CR test should include. Their recommendations included: (1) use as many items as are necessary to demonstrate that the student can perform under specified conditions, selecting the objects and conditions with which the student must work and (2) use as many items as are necessary to ascertain that the criterion objectives have been met.

In order to guard against chance occurrences during the test. Sweezey and Pearlistein indicated that more than one of the same type of item might be included. For example, in a typing test there might be two equivalent passages to be typed, within the same time constraints, but on two different types of manual typewriters. There are situations in which a single occasion of performance may not be an accurate indication of complete mastery of the task. Gagne (1969) stated that two samples of performance of a single class (i.e., two items) should suffice to test whether a student can or cannot perform that class of behaviors. One item was not considered sufficient because there may well be unknown factors influencing the responses to any given item. However, Gagne believed it unlikely that these unknown factors would operate in the same fashion in the case of two items. Thus, according to Gagne, a test should include two items of each type; performing correctly either item is taken as a demonstration of mastery of that objective.

Summary

As was the case with cut-off score determination, the choice of test length depends on many factors. The variables to be considered range from testing time, convenience, and fatigue to test reliability, and validity. It seems that where merited by the situation, a cost-benefit analysis needs to be completed. Again, there was little convergence in the literature relative to a "preferred" approach.

V. CR PERFORMANCE TESTING AND RATER ERROR

One specific type of CR test is based on performance demonstration and is particularly relevent to technical training in a wide variety of Air Force schools. Generally, but not exclusively, this type of test employs a checklist rating for format. The present chapter discusses performance rating procedures in the armed services, possible causes of rater error in such checklists, and a general model of rater behavior.

Shriver and Foley (1974) summarized the advantages of CR performance tests:

> Paper and pencil job knowledge tests are more easily developed and administered...They require no equipment. They usually require less time to administer. But they do not measure how well individuals can perform the tasks (which they are learning to perform and for which they eventually will be paid to perform) ...Such job knowledge tests are no bargain, no matter how cheaply they can be developed or how conveniently and easily they can be administered(p. 58).

Shriver and Foley held that paper and pencil tests are not empirically valid for measuring job ability. Convenience of administration, ease of scoring, ease of interpretation, tradition, and lower costs were the explanations given for the continued use and reliance on the written tests.

Performance Checklists

The performance checklist is employed as a device for measuring student ability when measurement of performance in process is of interest. This type of measurement is required when no measurable end product is involved, when sequence of performance or adherence to prescribed procedures is important, or when subsequent steps might cancel errors made in prior steps. In some cases, both performance in process and the final product are scored. The product may be examined for accuracy, freedom from defects, and ability to meet operational requirements. Siegel (1971) indicated an important problem inherent in the checklist, i.e., certain aspects of a job may be lost in the checklist approach. For example, situations have been identified in which a performance score does not correlate highly with expert judgments of the quality of the final product. Where checklist scores do correlate highly with expert opinion of the final product, Siegel supported the checklist as the preferred instrument. The reasons offered to support this contention were that: (1) more objectivity may be incorporated into checklists than in many other instruments, (2) increased inter and intraexaminer reliability, (3) increased test reliability, (4) less examiner experience in the particular task is required, and (5) the checklist is a valuable diagnostic tool allowing insights to be gained by the examiner as the student performs the task.

Sweezey and Pearlstein (1975) stated that the checklist is the most "reliable rating scale.". They maintained that the format of the checklist, with its emphasis on elements of behavior, assists the evaluator in his evaluations of each performance step and that the checklist reduces the effects of many of the errors found in ratings because of its minimization of subjectivity. Unlike the typical rating scale, checklist content and scoring do not deal in trait generalities but emphasize specific observable behaviors. The checklist asks whether or not specific behaviors are demonstrated. It is a list of behavioral statements on which check marks are place only for those behaviors that are demonstrated.

Performance Checklist Development

The performance checklist is usually developed from a task analysis. Guion (1965) reported use of the critical incident approach to developing checklists. A critical incident analysis describes those behaviors that are "critical" to successful task performance. Critical incidents are such that, depending on whether or not the "critical" behaviors are performed, success or failure will result. No matter what approach is used, the major result is that a list of specific, behaviorally based task statements are produced.

Performance Checklist Scoring

Checklist items can be scored in several ways. Typically the checklist is just that--a list of behavioral statements along with a space for entering a check mark if the behavior is performed. This would result in a one or zero score for each item of behavior. Scores may be derived from such checklists much as for tests in general. Scoring may be accomplished by summing across subscores of the inclusive components of the global task. In addition, criteria of mastery/nonmastery may be established from total scores.

Another common scoring method, the method of summated ratings, allows the rater several response categories for each behavioral item. These response categories usually follow a Likert-type format, such as "strongly agree," "agree," "undecided," "disagree," and "strongly disagree." Here the rater judges the amount that he agrees or disagrees that each listed behavior describes the examinee's performance. Each response category is numerically weighted where, for example, given a desirable behavior, weights might range from 5 for the "strongly agree" response to 1 for the "strongly disagree" response. An overall rating for performance on the task can be accomplished by simply summing the response weights across items. A variant of this approach allows the rater to assign from zero to some given point--depending on performance quality--for each listed behavior.

Examples of Performance Checklists in the Navy and Army

All the military services employ some form of checklist for performance assessment. In the Navy, Abrams and Pickering (1962) identified four characteristics that should be built into a checklist. They contend that a checklist should:

1. possess ease of administration, scoring, and interpretation
2. evaluate proficiency in essential areas of performance
3. point to the essential areas of training if proficiency is lacking
4. impose time demands (if appropriate)

Abrams and Pickering developed a checklist for measuring the maintenance proficiency of naval sonarmen in the fleet. This checklist required supervisors to evaluate sonarmen performing routine checks. They marked Y(yes) or N(no) alongside each step in the checklist. The pupose of the checklist was to identify areas where additional training was required. The N scores indicated areas of weakness and, perhaps a requirement for training.

The Army uses the Skill Qualification Test (SQT) to evaluate an enlisted soldier's performance on specific tasks (SQT - A guide for leaders, (1977). These tests measure a soldier's skill in performing tasks. The SQT involves an evaluation of critical tasks. A soldier is proficient if he reaches or surpasses a fixed standard. The SQT has two components--written and hands-on. The hands-on component is the performance aspect of the SQT. The observations are made by the supervisor as the soldier performs various tasks. The SQTs are administered to the soldier every two years. The results are used as qualifiers for advancement to higher skill levels. The tasks are listed in checklist fashion. Scoring is performed by the Pass-Fail technique. (Department of the Army Scoring Booklet, 1977).

Other uses of the checklist in the military context are found in: Richlin, Federman, & Siegel, 1958; Siegel, Richlin, & Federman, 1958; Richlin, Siegel, Schultz, & Benson, 1960; Siegel, Richlin & Federman, 1960; Siegel & Schultz, 1960; Schultz & Siegel, 1961; Siegel, Schultz, Fischl & Lanterman, 1968.

Performance Checklists in the Air Force

In order to examine the utilization of criterion-referenced measurement in the Air Force, a sampling of performance checklists in Air Training Command resident courses was taken. After a comprehensive selection process (designed to cover professional, technical, and clerical occupations), eleven courses at two Air Force bases were selected and certain checklists from each course were identified for intensive study.

Generally, the results of this study indicated the following: (1) The instructors based performance checks on the criterion objectives. However, in only a few instances were the standards of speed and/or accuracy specified in the criterion objective. (2) In most situations, the procedural steps (the checklist) involved in the performance checks were taken from technical orders and/or the student workbooks. Although acquisition and retention of skills was a specified goal of the performance checks, the performance checks as conducted could not satisfy this goal. There was little evidence of criterion referencing in the formal sense. The conditions for successful performance were not clearly specified and thereby decreased the overall effectiveness of the check. (3) The performance check is an appraisal of student's ability to perform tasks on which they received training. In most instances, familiarity with the equipment and the tasks involved in the training received was the only requirement for passing the performance check. For these situations, proficiency demonstration was not a requirement. The instructor or the students' peers were often allowed to demonstrate or prompt the student during the performance check. Accordingly, standardization of testing procedures was lacking in these areas. (4) The courses examined in this survey indicated a ratio of performance checks to training objectives ranging from 32 percent to 98 percent. In most cases, if a performance check could not be accomplished, a training deviation was required.

Until about 10 years ago, performance tests (as opposed to performance checks) were sometimes administered in Air Force courses. The results of these tests were used in conjunction with the results of written tests (used for testing knowledges rather than skills) to develop an end-of-course grade for each student. Performance testing required that students first be provided training time on the equipment and then brought back for a final evaluation. It was said that this procedure was unsatisfactory because: (1) It tied up equipment needed for other purposes, (2) Test materials consumption was costly, and (3) Test administration time (if students were to be tested on all tasks taught in the course) was excessive. While such performance tests are no longer administered, the reasons for their disuse might be reexamined, as will be suggested in Chapter VIII of this report.

Current training programs are designed to provide students with a 3-skill level of training. On-the-job training is intended to pick up where the technical training program left off. However, postcourse training may be costly. Such costs are not considered within the usual training cost need for more thorough performance testing.

Rater Error

Performance checklists are not without weakness. They are subject to rater error. The accuracy of a rating derived from a performance checklist rests on the assumption that the, "Human observer is a good instrument of quantitative observation, that he is capable of precision, and some degree of objectivity." (Guilford, 1954 p. 278). Unfortunately, the human observer is not always capable of either precise or objective judgments. Barrett (1966) wrote: "The ideal rater, who observes and evaluates what is important and reports his judgments without bias or appreciable error, does not exist, or if he does, no one knows how to distinguish him from his less talented colleagues." (p. 99).

For human observations to be used successfully, the error inherent in ratings must be reduced. Reducing such errors is a multifold task. Raters who possess the knowledge necessary for making accurate judgments must be located and trained. The reporting format must be structured so that the raters can make judgments which possess minimum errors. Errors can also result from biases within the rater. The rater may have had prior association with the examinee and opinions may have been formed which are incidental to the structured task on hand.

While the structured checklist format attempts to reduce such bias, a number of types of rating error remain as potential problems even in the structured performance test checklist context.

Systematic and Random Error

Theoretically, for any kind of measurement, two types of erro can occur: (1) systematic, and (2) random. Systematic error is called bias and is constant across any measurement taken with a specific measurement instrument. For example, a bias of .15-inch multiples can occur if a 12-inch ruler is actually 12.15 inches long. Random error, as the name implies, occurs unsystematically across the measures taken and the measurement instruments used. Random error is independent of the specific measurement instrument and may be assumed to average out across repeated measurements.

With ratings, systematic bias is typically called "rater bias" (as the rater is the measurement instrument) and random error is called "error."

General Model Of Rater Behavior

A general model of rater behavior can be defined, which considers the two types of rater error. Given a rater and his rating of the performance of an individual, the observed rating score consists of the examinee's true performance plus the bias and random error that occurs:

$$X_O = X_T + X_B + X_E$$

where:

X_O = the observed rating
X_T = true score
X_B = bias
X_E = random error.

Expanding the model across raters and ratees, raters behavior can be described by the components of variance accounted for in the observed ratings:

$$\sigma_O^2 = \sigma_T^2 + \sigma_B^2 + \sigma_E^2$$

where:

σ_O^2 = the total rating variance observed
σ_T^2 = the variance due to true scores
σ_B^2 = the variance due to bias
σ_E^2 = the variance due to random error.

Rater Bias

In concept, random error is a catch-all category into which all unaccounted for or unexplained variance falls. Random error is the residual variance component left after the true and bias variance components have been subtracted from the total variance. Bias, on the other hand, has specific conceptual and operational definitions. Several sources of bias have been defined. Guilford (1954) defined the following systematic biases of the individual rater.

- Leniency Bias. Leniency exists when the rater choses to rate individuals very leniently (or very harshly). Leniency may occur because the rater has some interest in the person being rated or because he is an "easy rater." Assuming that the examinees are normally distributed with respect to the rated variable, a leniency bias is manifested by a distribution of ratings that is skewed to the left or to the right.

- Central Tendency Bias. Central tendency bias exists when a rater avoids making extreme judgments. Instead of using the high and low ends of a rating scale, as well as the central area, the rater tends to group ratings around the central area of the scale. Such a general tendency results in an artifical restriction of the measurement range.

- Halo Bias. Halo bias results when a rater systematically rates a person too high or too low on all items under consideration. This type of bias is called the "halo effect" and is quite common. The halo effect is believed to stem from an overall impression (e.g., a favorable or unfavorable impression) a rater holds about the ratee. The rater generalizes this overall impression to the specific items under current evaluation. This error results in spurious positive correlations between rated items.

- Logical Bias. Logical bias exists when there is a systematic tendency to rate in the same manner traits that appear to be related. For example, a rater who rates an individual high on tool use may tend to rate the individual high on care of tools. Logical bias is manifested by intercorrelations between items that in fact are not interrelated.

- Contrast Bias. Contrast bias is the tendency on the part of the rater to rate others lower than himself on a given item or set of items. For example, the instructor who considers himself high on a given performance may tend to rate students as lower than himself on the item. Similarly, the rater who views himself as exceptionally low in the performance may tend to view others as high. This error is seen as the interaction between performance items and raters.

- Proximity Bias. Proximity bias results from the tendency of a rater to rate similarly items which are close together on the rating form. This type of error is also known as the order-effect. Thus, the order that the items are rated may influence a rater's judgment. The error of proximity is shown in spuriously positive correlations between adjacent items on the rating form.

There are, of course, other biases possible. For example, a bias will result if the rater rates dishonestly. There may be raters who are openly dishonest or hostile to the rating procedure or individual being evaluated. Alternatively, the rater may be uncommitted to the rating task. Campbell, Dunnette, Lawler, and Weick (1970) suggested that lack of rater commitment is the most serious source of rating bias. They reasoned:

> The most serious source of difficulty (bias) is a very fundamental one - stemming from a common tendency for psychologists to impose their own beliefs about job behavior and their own systems for recording it upon the persons whose task it is to observe that behavior...(It is) a lack of understanding and a lack of commitment to the observational (rating) task on the part of the observers. As a consequence, they (the observers) tend to fill in the forms (job behavior rating scales) with little conviction; the records contain large and for the most part inestimable error, (p. 118-119).

Minimization of Rating Error

Given the above definitions of rating errors, it becomes obvious that the major purpose of a checklist procedure is to minimize bias and random error. There are several widely used techniques which attempt to minimize error in the job performance rating situations. These involve such methods as carefully planned form construction and presentation techniques and involved scoring methods. Random presentation of items to be rated and random inversion of the rating scales has been suggested, and is often used. Training the raters to make them aware of the potential pitfalls has also often been suggested.

Statistical Models of Rater Behavior

More relevant to the present purposes are models of rater behavior which evaluate rating data for the presence of rating errors and attempt to correct for such errors statistically. These models conceptualize and operationalize bias and random error. Typically, the models are specifications of the general variance model of rater behavior presented earlier in this chapter. By definition, such models are linear and follow the analysis of variance paradigm. Guilford (1954) presented one model of rater behavior. Winer (1971) and Cronbach, Gleser, Nanda, and Rajaratnam (1972) also suggested variance analytic approaches.

Guilford's Model of Rater Behavior

Guilford (1954) extended the genral linear model of rater behavior by defining the components of rating scores across raters, traits, and ratees. The definitions are:

X_{ijk} = a rating of ratee I on trait J by rater K

X_{ijt} = the "true" score of ratee I on trait J

X_{ijke} = the total error in rating X_{ijk}

and:

$$X_{ijk} = X_{ijt} + X_{ijke}$$

Thus, the rating provided by a rater is modeled as the linear combination of two components, truth and error. From here, Guilford further elaborated on the model. The total error X_{ijke} was broken into four separate and additive rating error components.

As noted, rated error consists of two types, bias and random error. In his model, Guilford (1954) operationalized bias and random error with these definitions:

X_{kl} = leniency bias; rater K's tendency to over or under value ratees in general

X_{ki} = halo effect; rater K's general tendency to over or undervalue ratee I across traits

X_{kj} = contrast bias; rater K's tendency to generally over or undervalue a certain trait across ratees

X_{ijkr} = random error; residual error made by rater K in rating ratee I that includes everything in X_{ijke} not otherwise identified

where:

$$X_{ijke} = X_{kl} + X_{ki} + X_{kj} + X_{ijkr}$$

and hence:

$$X_{ijk} = X_{ijt} + X_{kl} + X_{ki} + X_{kj} + X_{ijkr}$$

In short, Guilford attempted to explain rater behavior in terms of true ratings, plus leniency, halo, and contrast bias, plus the undefined random error. Other biases, such as central tendency, logical, and proximity, were left undefined because these errors are nonincremental, nonadditive biases and thereby do not fit the linear model. If they occur, such biases, by operational definition are grouped together with random error.

In concept, rating error occurs whenever rating variance results from: (1) the sole effect of the raters or, (2) the interaction effects of the raters with ratees and traits. This follows because the rater per se should have no effect on any given rating. The rating should accurately reflect the ratee's true position on the item rated. Any effect by the rater on a ratee's rating on the item rated is spurious resulting from bias and random error.

True rating variance, on the other hand, occurs when rating variance results from differences in examinees and rated variables. Conceptually, this follows because examinees can differ on their true positions on the rated traits. It is fact that individuals can differ and that these differences would occur with respect to performance items. Note, however, true rating variance does not necessarily accurately reflect the examinee's true positions on the rated variables. In a sense, true does not mean truth. True rating variance is truth to the extent the examinee and rated item differences exist and are rated as such. This extent can not be measured. Operationally, the truth is measured by the variance resulting from the average rating across raters for given examinees and rated items. A rating which is averaged across raters is a better estimate of the examinee's true position but the possibility of a poor estimate remains. In all, true rating variance is so distinguished because the probability of accurate, true rating variance occurring in these variance components is greater than that in the defined error rating variance components.

Correction for Rater Bias

A final consideration in Guilford's (1954) model is the correction for bias in ratings. In his model, ratings are corrected for bias by subtracting the bias variance components from the total rating variance. With a three-way variance analytic design, the correction for bias is defined by:

$$X'_{ijk} = X_{ijk} - (C_k + AC_{ik} + BC_{jk})$$

where:

X_{ijk} = a rating of examinee I on item J by rater K

C_k = rater's leniency bias

AC_{ik} = rater's halo bias

BC_{jk} = rater's contrast bias

Hence, rating data are adjusted and left composed only of true variance and random error variance. The rater's leniency, halo, and contrast biases are statistically eliminated from the rating data.

According to Guilford (1954), the results of bias corrections in rating data are such that:

> The variance remaining in such values would be made up to a larger degree of the true-value contribution. Since reliability of measures is defined as the proportion of true variance in them, the ratings should then be more reliable and their possibility of correlating with other measures should be increased. Hence, there would also be the possibility of increased validity (pp. 282-283).

and

> What effects should the adjustments have upon the correlations of the ratings? We have the possibility of computing rater intercorrelations, which indicate the internal consistency among raters. Such correlations have usually been regarded as indices of rating reliability but sometimes as rating validity (pp. 286-287).

Extension of Guilford's Bias Correction

To further describe and extend Guilford's (1954) model, a three-way variance analysis--ratee by trait by rater--without replication design may be employed. An example of such a design was presented by Cronbach, Gleser, Nanda, and Rajaratnam (1972). Here, rater behavior is explained by seven separate components of rating variance. A possible eighth component--the population rating value--is defined as zero and thereby eliminated from the model. The variance form is:

$$X_{ijk} = A_i + B_j + C_k + AB_{ij} + AC_{ik} + BC_{jk} + ABC_{ijk}$$

where:

- A_i = effect due to ratees being rated differently
- B_j = effect due to traits being rated differently
- C_k = effect due to raters rating differently
- AB_{ij} = effect due to ratees being rated differently on the separate traits
- AC_{ik} = effect due to ratees being rated differently by the individual raters
- BC_{jk} = effect due to traits being rated differently by the individual raters
- ABC_{ijk} = residual rating values.

Chapter VI of this report describes the methods, procedures, and results of an empirical study which tests the utility of Guilford's model and the resultant corrections for bias in the Air Force technical training context.

VI. RATER BIAS AND ITS CORRECTION--EXPERIMENTAL STUDY

One theoretical view considers CR measurement (e.g., Gagne, 1969; Ivens, 1970; Shriver & Foley, 1974) and performance measurement to be distinct and unrelated. A second view of CR measurement (e.g., Crehan, 1974; Glaser & Cox, 1968) considers these two topics to be intimately related due to the emphasis of both on performance. In practice, CR measurement often consists largely or completely of performance testing. As pointed out earlier, several types of bias typically weaken rated measures of performance. The present chapter deals with an experimental investigation of the adequacy of the scoring of a type of CR performance rating, the performance checklist, employed by the Air Force Technical Training Schools.

Specifically, rater bias in the case of a current performance checklist was investigated along with the utility of a statistical model for correcting such bias. Although the statistical model was developed earlier by Guilford (1954), its usefulness in the performance test sphere has, to our knowledge, not been previously investigated.

Method

The evaluation of bias in checklist employment was based on the procedures originally developed by Guilford (1954) and expanded by Cronbach et al. (1972). Guilford presented a model for evaluating a set of data for rater bias and for correcting the data set so that the effects of rater bias can be accounted for. To obtain the required data, one performance checklist used in a selected Air Force course (referred to as Course A) was employed. The students in the course were rated as teams, and for this reason teams are considered the unit of measurement in the analysis. Such an approach parallels one suggested by Campbell and Stanley (1963, p. 23).

Analysis of the data was undertaken at two levels of granularity. A satisfactory (S)-unsatisfactory (U) dichotomy formed the basis of one set of analyses. This analysis has face validity as the course employs the test scores in just this manner. An analysis at a second, finer level of granularity was also undertaken.

This second analysis considered exact numerical scores (prior to categorization as S or U). This second analysis was expected to provide a more sensitive and accurate appraisal of the reliability and bias of the performance checklist measurement process.

Sample

A total of six, three-man student teams from Course A were rated by four instructors/evaluators. The students were regular members of several classes of Course A at Lowry AFB. The students included both enlisted men and officers. This composition was typical of a Course A class.

The students sampled were organized into teams. These teams, rather than individual students, were the basic units rated. All performance checks took place at the completion of the course.

Raters

Four instructional staff members of Course A served as raters. The relevant experience and present job of the raters is given in Table 6.1. The raters acted independently in making their ratings.

Table 6.1

Background of Raters for Course A

Ratee Designation	No. of Years in Service	Exper. In Area (Years)	Exper. Teaching Course (Years)	Total Teaching Exper. (Years)
N	9.75	8.75	2.75	2.75
I	15.5	8.5	.08	9.75
R	12.0	8.0	1.75	*
T	17.0	8.0	1.0	1.0

*Data not available.

The performance check was based on a field problem, one of several team exercises given at the end of the course. For present purposes, the exercise was shortened so that it would take about three hours for completion. Performance check scoring was performed through the use of the Student Progress Checklist from the course. The Student Progress Checklist from the course covered four major performance areas. The rating procedure produces an overall score which is dichotomized, in practice, into pass and fail categories. The field problem situation was prepared in a standard fashion and introduced using a standard introductiory message.

Introductory

The students were all given standard instructions which asked them to perform the tasks they had learned, while paying full attention to safety. Directions concerning available tools, resources, and verbal access to the instructors were also included. It was anticipated that the students might feel ill at ease in the unorthodox situation to follow, in which several instructors would be observing the student's performance. The instructions attempted to counter this before it developed and to establish a nonthreatening atmosphere.

Rater Training

A detailed integrated briefing was given to the instructors prior to the first day of testing. This briefing had three main objectives: (1) to assure that the instructors knew their specific roles and duties, (2) to assure that the evaluation would not be construed by the instructors as threatening to them as individuals or as members of the school's instructional staff, and thereby, (3) to assure that the instructors were motivated to cooperate and help perform in accordance with instructions. Relevant to the duties of the instructors, it was stressed that the ratings should be made as they are typically completed during the testing process in the course. Several rules specific to this test administration were given. These were designed to assure independence of the raters in making their evaluations.

Control Over Conditions

An Applied Psychological Services' staff member supervised all aspects of the data collection. Care was taken to ensure that there was no opportunity for students, already rated, to converse with students yet to be rated. The actions of the raters were also closely observed. When necessary, the instructors were reminded that it was important that their ratings be made independently, and that their score checklists be kept confidential. Due to the intrinsic motivation of the raters and the close supervision given, the integrity of the data collection process was assured. During the course of the evaluations, the instructors were quite conscious of their responsibility to give independent ratings and to refrain from giving feedback to either the students or to the other raters concerning their judgments of the ratees' performance.

Results

The effects of several important sources of rater bias on the quality of ratings observed in Air Force technical schools can be evaluated and quantified through methods first suggested by Guilford (1954). To our knowledge, they have not been previously employed in the performance check situation. Following Guilford, the data were conceptualized as a three way factorial design (rater by trait by ratee). In this situation, Guilford identified three sources of rater error in terms of main effects and interactions. Leniency

errors are manifested in the rater main effect. Halo effects are revealed in the rater-ratee interaction. (Recall, ratees here are teams). And finally, contrast error (operationalized as the tendency of individual raters to over or undervalue certain traits) is manifested in the rater-trait interaction. These relationships are given in Table 6.2.

Table 6.2

Components of Rater Bias

Bias Component	Corresponding ANOVA Effect	Component Symbol
leniency	rater effect	X_{kl}
halo	rater-ratee interaction	X_{ki}
contrast	rater-trait interaction	X_{kj}
random error	residual	X_{ikjr}

Variance Analysis

In the variance analyses that follow, the effects of three measurable sources of rating variance due to rater bias are tested for significance. Then, those statistically significant variance estimates are quantified and the original data corrected for bias based on these estimates of the size of the bias. Finally, an evaluation of the effectiveness of these corrections is conducted.

The performance checklists were completed by the instructors according to the process currently in use by the instructional staff of the course. For each team, four checklist areas were scored, which we shall designate Part One, Part Two, Part Three, and Part Four. In the variance analytic design, these four performance areas are called traits so as to be consistent with Guilford's terminology. For each of these four traits, both dichotomous S and U scores and continuous numerical scores were obtained. The S and U scores were based on an overall rating of a team's performance on a given trait by the raters. The numerical scores were based on the sum of points awarded the team (from a total possible number of points) by the rater from its performance of the individual task items within a given trait. As the total number of points possible varied from trait to trait, the numerical scores were converted to percentage scores for comparison purposes. The S/U scores and the percentage scores formed the bases for all data analyses. An example of how one trait, Part One, was broken into individual items is presented in Exhibit 6.1.

Exhibit 6.1

Fragmentation of Part One into Task Items

I. Part One	Possible Points Awarded by Rater	
a. T.O.		
(1) ....?	6	
(2) ....?	6	
(3) ....?	4	
(4) ....?	6	
Sum		22
b. T.O.		
(1) ....?	6	
(2) ....?	6	
(3) ....?	3	
(4) ....?	6	
(5) ....?	6	
(6) ....?	6	
Sum		33
c. Mission		
(1) ....?	10	
(2) ....?	5	
(3) ....?	10	
(4) ....?	2	
(5) ....?	10	
(6) ....?	6	
(7) ....?	2	
Sum		45
Total		100

Though apparently violating the interval scale of measurement requirement, the use of analysis of variance for dichotomous data is acceptable. Winer (1971, pp. 293-296) provided examples of variance analytic designs with dichtomous data. And, Lunney (1970) presented evidence for the robustness of the approach with dichotomous data.

Leniency

In theory, rater bias refers to any systematic departure from a true score. The true score of a student is, in the situations we dealt with, beyond our measurement capability. Therefore, any analysis of rater bias must consider the errors of a rater as compared to all of the raters employed in the study. Consider leniency bias as a case in point. The amount which raters deviate from the theoretical "true" rating is indeterminable. An estimate of overall rater leniency error which substitutes for the unobtainable theoretical value is derived which is based on the agreement of the four raters evaluated. For both data sets, the statistical significance of this effect can be tested using a variance analytic test of the main effects of raters.

For the S and U ratings, coded 1 and 0 respectively, the F test (Table 6.3) for the interrater effect showed the rater leniency effect as insignificant, $F(3,45)= 1.07$, $p<.05$. Conversely, for the numerical data, the interrater effect was statistically significant, $F(3,45)=6.45$, $p <.01$ (Table 6.4). Accordingly, a leniency bias was evidenced for the numerical data.

Halo

The halo effect is defined here as a spurious over or under estimation of a specific team by a specific rater, or by group of raters (but not by all raters). As before, the halo error of a rater can only be evaluated in terms of the ratings of the other raters. Specifically, it is evaluated using the variance analytic test of the significance of the rater by team interaction.

For the dichotomous data, as indicated in Table 6.3, the test was not statistically significant ($F(5,45)=.71$, $p <.05$); for the numerical data (Table 6.4), the team by rater effect was statistically significant ($F(15,45)=3.15$, $p <.01$). For the numerical ratings then, the results suggested that a halo effect contributes a reliable source of error to the observed ratings. Furthermore, this team by rater variance estimate probably reflects a conservative estimate of the actual halo error occurring. Because the team by rater variance estimate is based on ratings averaged across traits, a halo effect in which one rater's evaluation of a team's trait affects the other trait ratings, but not all to the same degree, is not included in the estimate. This incomplete type of halo effect would, in the present type of analysis, become a part of residual error variance.

Table 6.3

Summary of Analysis of Variance of Satisfactory-Unsatisfactory Ratings for Six Teams on Four Traits by Four Raters

Source of Variance	Sum of Squares	Degrees of Freedom	Mean Square	F
Raters (R)[a]	.365	3	.122	1.073
Traits (T)	5.114	3	1.705	15.060**
Teams (I)	1.594	5	.319	2.815*
RXT[b]	2.094	9	.233	2.056
RXI[c]	1.198	15	.080	.706
TXI	3.948	15	.263	2.325*
residual	5.094	45	.113	
total	19.407	95		

[a]Reflects leniency bias
[b]Reflects contrast bias
[c]Reflects halo bias

**p < .01
*p < .05

Table 6.4

Summary of Analysis of Variance of Numerical Ratings for Six Teams of Four Traits by Four Raters

Source of Variance	Sum of Squares	Degrees of Freedom	Mean Square	F
Raters (R)[a]	3,248.9	3	1,083.0	6.47**
Traits (T)	6,570.9	3	2,190.3	13.08**
Teams (I)	2,361.6	5	472.3	2.82*
RXT[b]	4,515.8	9	501.8	3.00**
RXI[c]	7,921.1	15	528.1	3.15*
TXI	3,579.4	15	238.6	1.42
residual	7,536.8	45	167.5	
total	35,734.5	95		

[a]Reflects leniency bias
[b]Reflects contrast bias
[c]Reflects halo bias

**p < .01
*p < .05

Contrast

The third and final source of rater bias to be considered is the contrast error. In testing the rater by trait effect, it was found that the effect was statistically significant (Table 6.3) for the numerical data ($F(9,45) = 3.00; p < .01$) but not (Table 6.4) for the dichotomous data ($F(9,45) = .08; p > .05$). Accordingly, it appears that contrast bias contributes a reliable source of variation to the numerical but not the dichotomous ratings.

In summary, the variance analytic evaluation of these three sources of rating error--leniency, halo, and contrast bias--suggested that all three were present in the numerical but not the dichotomous ratings. That is, the results indicated that the performance check ratings were more confounded with rater biases when individual task items were rated and summed to provide a trait score than when overall performance (on the individual traits) was judged as either satisfactory or unsatisfactory.

Discussion

Taken together, the results consistently show that the numerical ratings are more affected by bias than the S and the U ratings. Such results suggest that the raters did agree in overall performance evaluations, but they did not agree in the evaluation of performance on the individual tasks. It should be obvious that a rater main effect (leniency), a rater by team (halo) interaction, and rater by trait (contrast) interaction effect cannot be statistically significant unless the raters differ on their respective ratings. In short, where variance due to raters is high, then the agreement among the raters must be low and where variance due to raters is low, then the agreement among raters is high. Thus, rater error is the inverse of rating reliability.

As ascertained in the AF case study (Chapter V), it appears likely that the raters may often disagree on performance standards for individual task items. Without specified standards or definitions of success, the individual raters must develop their own standards for judging performance and these standards may be more subject to disagreement for the individual behaviors and performance than for an overall evaluation.

The reason why rater bias occurred notwithstanding, it is evident from the results that the numerical ratings were more subject to rater biases than the S-U ratings. This finding suggests the use of the overall S-U method over the numerical item scoring method, at least for the raters and the test situation here involved. However, such employment loses the diagnostic information offered by numerical item scoring.

One way to eliminate rater bias from item scoring results was suggested by Guilford (1954). In this method, the magnitude of rater bias--leniency, halo, and contrast--is estimated via the variance analytics design and these estimates are subtracted from the observed ratings so as to produce ratings with these effects removed. Such analyses were performed for the numerical ratings and are reported below. For reference, these analyses followed Guilford's (1954) procedures for estimating rater bias, correcting rater bias, and evaluating the corrections for rater bias.

Estimation of Magnitude of Rater Bias

Leniency Correction

Rater leniency error was estimated on the basis of the difference of the individual rater means from the grand mean. These means are presented graphically in Figure 6.2. For the four raters N, I, R, and T--these errors are +2.76, -9.01, +6.8, and -.55, respectively. In Guilford's notational system these errors are denoted X'_{kl}. The rater means and the rater leniency correction, X'_{kl}, are given in Table 6.5. The means are presented graphically in Figure 6.1.

Halo Correction

The magnitude of the rater by ratee (team) or halo error was estimated from the data by removing X'_{kl} and team effect (d_i) from the means over traits for raters and teams. The means over traits for raters and teams are shown in Figure 6.3 and Table 6.5A. The corrected means, which are an intermediate step in arriving at the correction scores, are given in Table 6.5B. The difference of these corrected means from the grand mean are estimates of the size of the halo effect for each rater-team combination. These range from -19.73 to +14.86 for the four raters. These 24 estimates are detailed in Table 6.5C. Following Guilford's notation these correction terms are denoted X'_{ki}.

Looking at the raters individually, rater I indicates the greatest degree of halo error. His ratings of teams 1 and 6 were the lowest of all teams as rated by all raters (see Figure 6.2). However, his ratings of the other teams were on average, typical of the ratings given by the other raters. Such a negative halo effect has been termed a "horns" effect; a plutonic reference. It is interesting to note that rater I was by far the newest instructor in the school. He was assigned to duty approximately two weeks before the data collection. As a point of conjecture, one might ask whether this rater's extreme tendency toward halo error will persist over time.

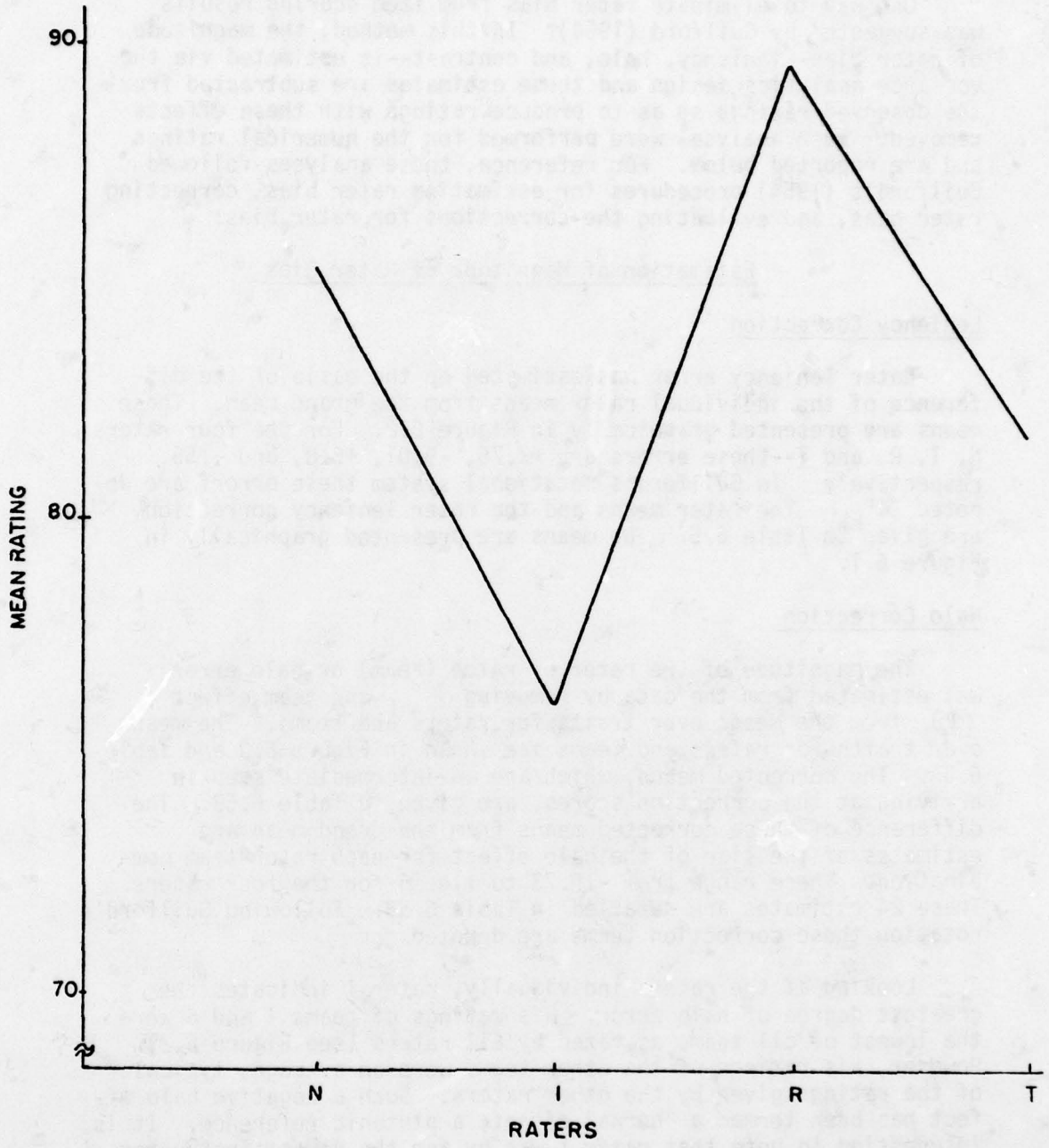


Figure 6.1. Numerical ratings of means over traits and teams by raters: leniency bias.

Table 6.5

Estimation of the Contribution to Rater Bias from Rater Effects (Leniency), and Rater-Ratee Interaction Effects (Halo)

A. Means over Traits by Teams and Raters

Rater	Teams 1	2	3	4	5	6	$\bar{X}_r$	X'kl (Leniency Bias)
N	83.63	86.70	93.23	75.93	87.23	84.60	85.22	+2.76
I	51.78	70.73	91.85	90.58	86.28	49.48	73.45	-9.01
R	78.30	86.68	98.03	94.65	90.78	87.13	89.26	+6.80
T	87.23	85.10	67.33	77.75	88.08	86.00	81.91	-.55
$\bar{X}_i$	75.23	82.30	87.61	84.73	88.09	76.80	82.46	
di	-7.23	-.16	+5.15	+2.27	+5.63	-5.66		

B. Means corrected for Rater errors (X'kl) and for Team deviations (di)

	1	2	3	4	5	6.	$\bar{X}_r'$
N	88.10	84.10	85.32	70.90	78.84	87.50	82.46
I	68.02	79.90	95.71	97.32	89.66	64.15	82.46
R	78.73	80.04	86.08	85.58	78.35	85.99	82.46
T	95.01	85.81	62.73	76.03	83.00	92.21	82.47
$\bar{X}_i$	82.47	82.46	82.46	82.46	82.46	82.46	

C. Contributions of Interactions of Rater & Ratee: Halo Errors X'ki

	1	2	3	4	5	6	Σ
N	5.64	1.64	2.86	-11.56	-3.62	5.04	.00
I	-14.44	-2.56	13.25	14.86	7.20	-18.31	.00
R	-3.73	-2.42	3.62	3.12	-4.11	3.53	.01
T	12.55	3.35	-19.73	-6.43	.54	9.75	.03
Σ	.02	.01	.00	-.01	.01	.01	.04

Notes:

1. $d_i = \bar{X}_i - \bar{\bar{X}}$; Team deviations from grand mean.
2. $\bar{X}_i'$ and $\bar{X}_r'$ are not equal to 82.46, the grand mean, due to rounding errors.
3. Summations (Σs) are nonzero due to rounding errors.

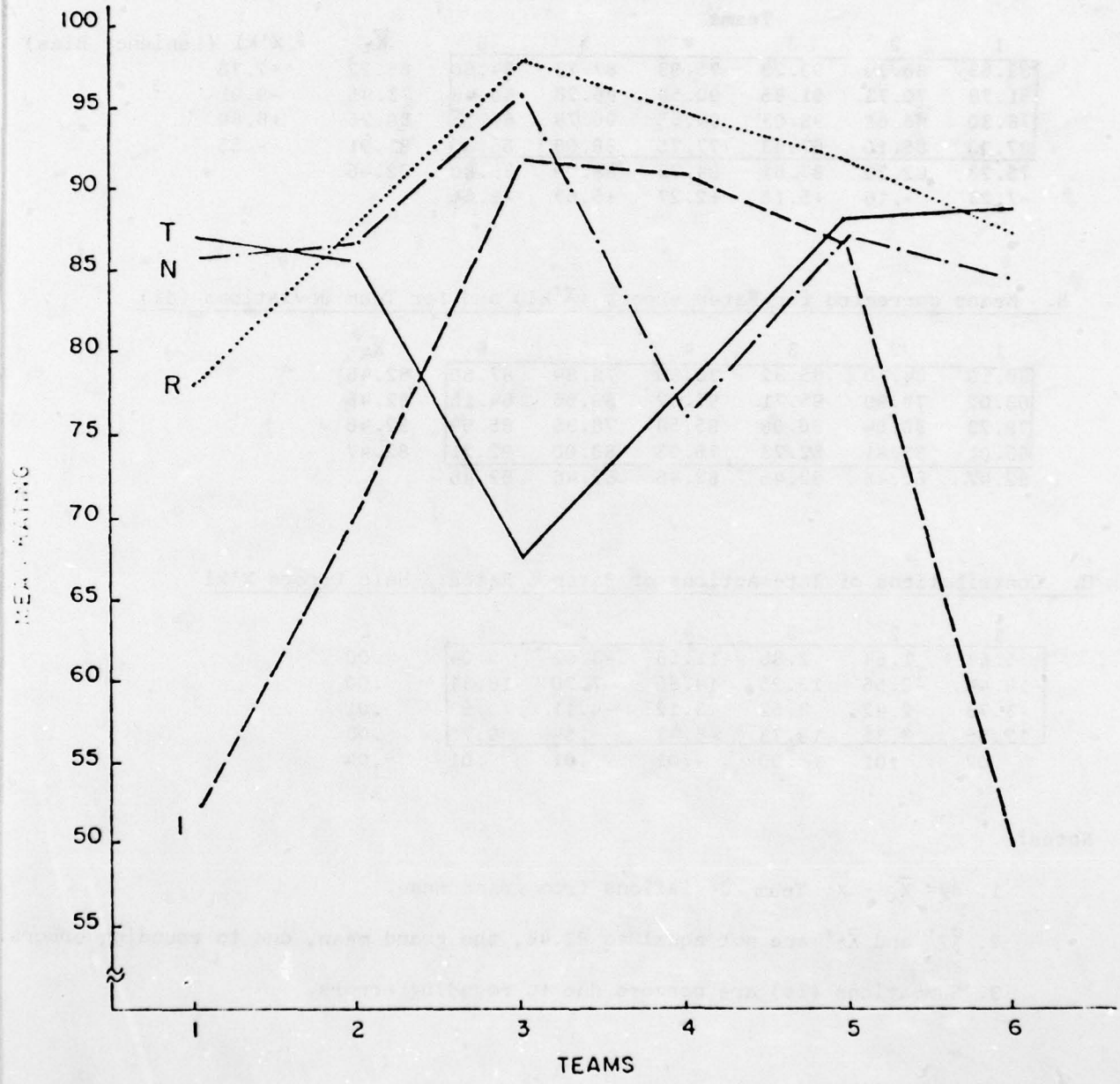


Figure 6.2. Numerical ratings of means over traits by teams and raters: halo effect.

Table 6.6

Estimation of the Contribution to Rater Error for Rater-Trait Interaction Effects (Contrast Bias)

A. Means over Teams by Traits and Raters

Rater	Traits P	M	D	C	$\bar{X}_r$	$\bar{X}'kl$ (Leniency Bias)
N	81.33	75.05	97.92	86.57	85.22	2.76
I	84.17	65.63	62.50	81.48	73.45	-9.01
R	88.57	75.78	100.00	92.58	89.26	6.80
T	92.33	56.17	87.50	91.65	81.91	-.55
$\bar{X}_t$	86.63	68.16	86.98	88.07	82.46	
dt	4.17	-14.30	4.52	5.61		

B. Means by Ratee & Trait, Corrected for Rater Error X'kl & for Trait Deviations dt.

	P	M	D	C	$\bar{X}_r'$
N	74.40	86.59	90.64	78.20	82.46
I	89.01	88.94	66.99	84.88	82.46
R	77.70	83.28	88.68	20.17	82.46
T	88.71	71.02	83.53	86.59	82.46
$\bar{X}t$	82.46	82.46	82.46	82.46	

C. Contributions of Interaction of Rater and Trait; Contrast Bias X'kj.

	P	M	D	C	Σ
N	-8.06	4.13	8.18	-4.26	-.01
I	6.55	6.48	-15.47	2.42	-.02
R	-4.76	.82	6.22	-2.29	-.01
T	6.25	-11.44	1.07	4.13	.01
Σ	-.00	-.01	.00	.00	-.03

Notes:

1. $dt = \bar{X}_t - \bar{\bar{X}}$; trait deviations from grand mean.

2. Summations ($\bar{\Sigma}$s) are nonzero due to rounding errors.

Contrast Correction

For an estimate of the size of rater contrast bias, the rater-trait interaction is considered. The magnitude of the contrast bias was estimated in a manner paralleling that described above for the halo effect. The X'_{kl} and the overall trait effects (d_t), were removed (subtracted) from the means over teams for raters and traits. The rating means over teams for raters and traits are shown graphically in Figure 6.3 and numerically in Table 6.6A. The corrected means are given in Table 6.6B. The differences between these corrected means and the grand mean are the estimates of the size of the contrast bias effect for each rater for each trait. These 16 estimates are given in Table 6.6C. They range in size from -15.47 to +8.18. Following Guilford's notation, these are denoted X'_{kj}.

Overall Correction For Rater Bias

Consider that each observed score, X_{ijk}, a rating of team i on trait j by rater k is composed of a true score (X_{ijt}), and the three sources of identifiable error considered above:

$$X_{ijk} = X_{ijt} + X_{kl} + X_{ki} + X_{kj} + X_{ijkr}$$

where X_{ijkr} is unexplained, residual error. We have derived estimates, X'_{kl}, X'_{ki}, and X'_{kj}, of X_{kl}, X_{ki}, and X_{kj}, respectively. It is possible to correct the X_{ijk} scores for these sources of error. This will theoretically yield scores (Y_{ijk}) that contain only residual error in addition to the true score. To accomplish this, each source of error was subtracted from each score, as follows. Let Y'_{ijk} be the estimate of $(X_{ijt} + X_{ijkr})$. Then,

$$Y'_{ijk} = X_{ijk} - X'_{kl} - X'_{ki} - X'_{kj}$$

The application of this process for statistically correcting rater bias can be visualized through an example. Consider the rating by rater N of team 2 on trait M (monitoring). This rating was 80.5. The rounded leniency bias, X'_{kl}, for rater N was +2.8 (see Table 6.5A). The rounded halo bias, X'_{ki}, for this rater and team was +1.6 (see Table 6.5C). And the contrast bias for this rater and trait, X'_{kj} was rounded to +4.1 (see Table 6.6C). The corrected score is therefore:

$$Y'_{N2M} = 80.5 - 2.8 - 1.6 - 4.1 = 72.0$$

Tables 6.7 and 6.8 present the original ratings and the ratings so corrected.

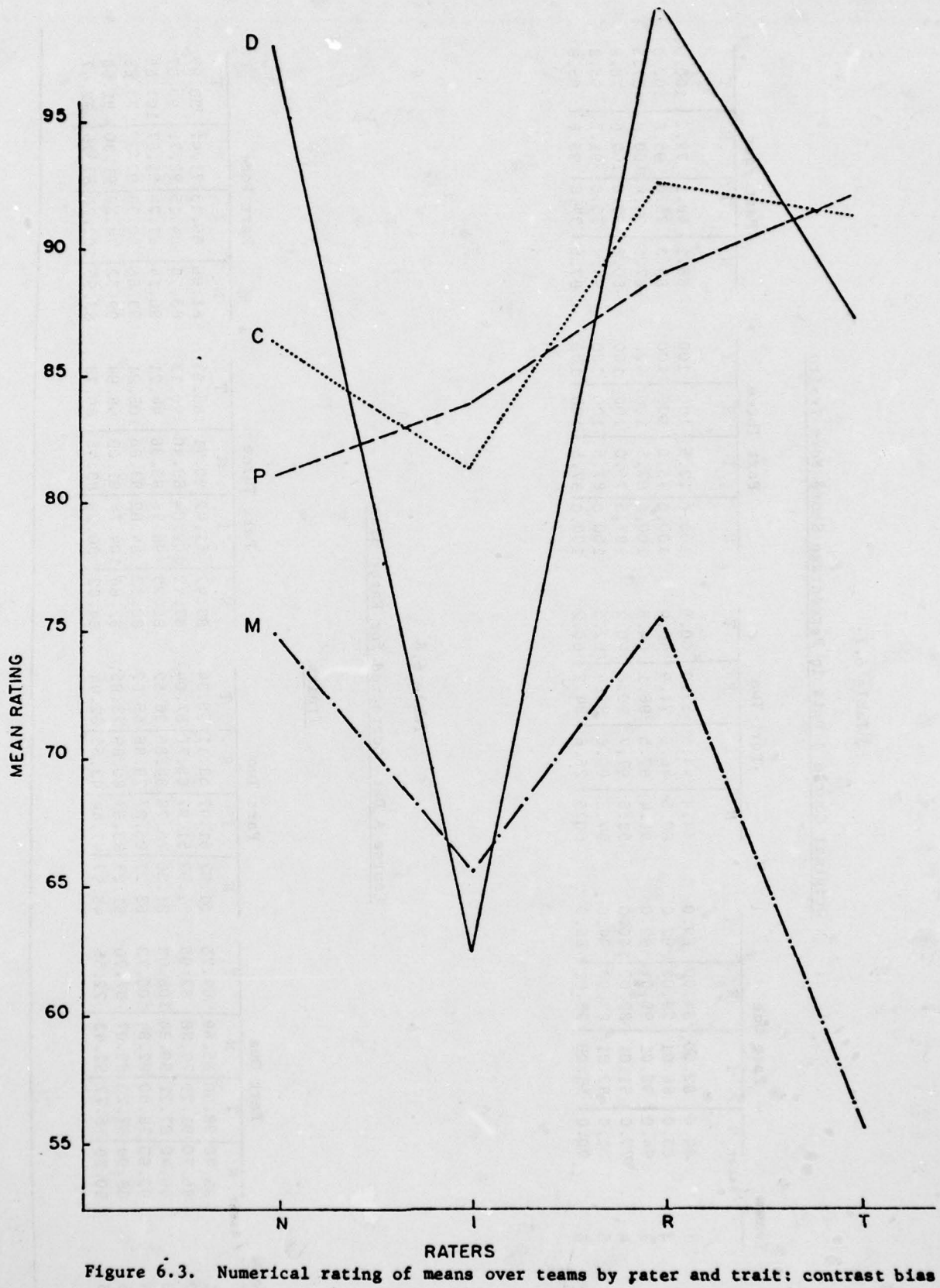


Figure 6.3. Numerical rating of means over teams by rater and trait: contrast bias

Table 6.7

Original Course A Data in Percentage Score Form Traits

Team / Rater	Part One				Part Two				Part Three				Part Four			
	N	I	R	T	N	I	R	T	N	I	R	T	N	I	R	T
1	84.0	82.00	84.00	88.0	64.4	44.5	57.0	60.9	100.0	12.5	100	100	86.1	68.1	72.2	100.0
2	83.0	86.01	79.00	92.0	80.5	46.9	71.9	48.4	100.0	75.0	100	100	83.3	75.0	95.8	100.0
3	94.0	98.01	96.01	92.0	91.4	87.5	96.1	64.8	100.0	87.5	100	25	87.5	94.4	100.0	87.5
4	77.0	91.01	88.00	100.0	58.6	97.7	90.6	40.2	87.5	75.0	100	100	80.6	98.6	100.0	70.8
5	60.0	92.01	87.00	94.0	94.5	90.6	84.4	62.5	100.0	87.5	100	100	94.4	75.0	91.7	95.8
6	90.0	56.00	98.01	88.0	60.9	26.6	54.7	60.2	100.0	37.5	100	100	87.5	77.8	95.8	95.8

Table 6.8

Course A Data Corrected for Rater Bias

Traits

Team / Rater	Part One				Part Two				Part Three				Part Four			
	N	I	R	T	N	I	R	T	N	I	R	T	N	I	R	T
1	83.70	98.90	85.69	69.75	51.87	61.47	53.11	60.34	83.42	51.42	90.71	86.93	81.96	89.13	71.42	83.87
2	86.70	91.20	79.38	82.95	71.97	51.97	66.70	57.04	87.42	102.04	89.40	96.13	83.16	84.15	93.71	93.07
3	96.40	87.21	90.34	106.03	81.65	76.78	84.86	96.52	86.20	98.73	83.38	44.21	86.14	87.74	91.87	103.65
4	93.90	78.60	82.84	100.73	63.27	85.37	79.86	58.62	88.12	84.62	83.86	105.91	93.66	90.33	92.37	73.65
5	68.90	87.26	89.07	87.76	91.23	85.93	80.89	73.95	92.68	104.78	91.09	98.94	99.52	74.39	91.30	91.68
6	90.30	76.77	92.43	72.55	48.97	47.44	43.55	62.44	84.02	80.29	83.45	89.73	83.96	102.70	87.76	82.47

Evaluation of the Correction for Bias

The efficacy of the corrections can be evaluated by considering: (1) the interrater agreement with respect to within trait judgments and (2) the intrarater agreement between traits. These are considered separately below.

By definition, rater bias decreases reliability through a reduction in interrater agreement. Removal of rater bias should therefore improve interrater agreement. To test this effect, the intraclass correlation (Winer, 1971, p. 283) among raters for each trait was calculated for the corrected and the uncorrected ratings. It was expected that the intraclass correlation among raters would be higher for the corrected as compared with the uncorrected ratings. Table 6.9 presents the intraclass correlation among raters by trait. As seen, all correlations increased: positive correlations became more positive while negative correlations became either less negative or positive. For one trait, Part Two, average rater agreement improved substantially.

The effect of statistically removing rater bias was further tested through examination of the bias introduced due to halo effect. Halo is expected to increase spuriously trait intercorrelations. Accordingly, if the statistical correction for rater bias was effective in removing the halo bias, then a decrease should occur in the intrarater between trait correlations.

The test of this hypothesis was based on Pearson product-moment correlation coefficients between all possible pairs of traits within raters calculated separately for corrected and uncorrected ratings. In all, 24 correlation coefficients were calculated on the corrected and 24 on the uncorrected data. Overall, the sought after decrease in intrarater agreement (with appropriate r to z transformation) was obtained. The mean of the correlation coefficients based on the uncorrected data was .34, while the mean for the correlation based on the corrected data was approximately zero.

Finally, the halo effect correction was examined through the intraclass correlation approach. Intraclass correlations among the traits per rater were calculated both for the corrected and the uncorrected ratings. Here, one would expect a decrease in the spuriously positive intertrait correlations produced by halo bias. Table 6.10 presents the resultant intraclass correlations. Again, the results suggested that, moving from the uncorrected to corrected ratings, the correlations moved to negative or more negative relationships. That is, the results indicated some evidence that the correction for bias was effective.

Table 6.9

Intraclass Correlation Among Raters by Trait for Corrected and Uncorrected Scores

Trait	Ratings Uncorrected	Ratings Corrected
One	-.08	.14
Two	.30	.70
Three	-.20	.09
Four	-.17	-.08

Table 6.10

Intraclass Correlations Among Traits for Halo Corrected and Halo Uncorrected Ratings

Rater	Halo Uncorrected	Halo Corrected
N	-.08	-.09
I	.48	-.17
R	.02	-.08
T	-.15	-.20

Variance Analytic Check

The variance analysis shown in Table 6.4 was reperformed employing the corrected (adjusted) data of Table 6.8. While the mathematics of the situation dictated that the leniency, contrast, and halo effects which were shown to be statistically significant (Table 6.4) should be zero as the result of the correction, such a recalculation provides an empirical check. Table 6.11 presents the results of this analysis. As anticipated, the three variance components of interest were reduced to zero with minimum, if any, effect on other variance components. This result lends further support to contentions that the statistical correction was effective in removing rater bias from the data.

Discussion

The results of the evaluation on the correction for rater biases provided general support to the contention that rater judgments, made during the use of performance checklists, can be improved through statistical methods. Given the small sample size in this study, general conclusions can certainly not be drawn. The correlations reported for both the uncorrected and the corrected ratings were low by any standard. This suggests considerable error in such ratings aside from that which was identified by the statistical techniques. Specifically, it seems that the residual (random) error in the numerical ratings may have reduced interrater agreement. Given the validity of Guilford's (1954) rater model, it is likely that the numerical ratings while possessing leniency, halo, and contrast bias also contained a large degree of random (unidentified) error. Some concept of the random error in the numerical ratings is shown in the sum of squares for the different sources of variance, (Table 6.4). The residual (unidentified, random error) sum of squares (7,536.8) is greater than 20% of the total sum of squares (35,734.5). Accordingly, while the adjusted ratings may constitute more true rating variance than the unadjusted ratings, it appears that the adjusted data still contained a large degree of random error.

If numerical ratings are to be used for valid decisions, they must be improved. If standards of performance are developed, and if the rater/instructors are trained to know and accept these performance standards, then agreement between the raters should increase. It appears from the case study (reported in Chapter V) that raters/instructors possess different standards of performance for individual task steps. Therefore, the low interrater agreement presently found might have been anticipated.

VII. SIGNAL DETECTION THEORETIC APPROACH TO ESTABLISHING RELIABILITY, VALIDITY, CUT-OFF SCORE AND UTILITY OF A CHECKLIST PREDICTOR EMPLOYED IN A TRAINING CONTEXT

The basic performance evluation method in the Air Force employs a behavioral checklist as the measurement tool. In such a checklist, the rater is provided with descriptive statements of task related behavior and he is asked to indicate those statements which are descriptive of the individual in question. Here, the rater is more a reporter of work behavior than an evaluator of performance. The results of the evaluation are employed to pass or fail a student or to classify the student as a "master" or "non-master." As stated earlier variability among student scores is often considerably reduced by this testing method. With reduced variability, the correlation procedures typically associated with norm referenced testing are difficult to employ for validation. In this chapter we suggest and describe a study into the possible use of signal detection theory for validating such checklists.

The chapter is arranged in three parts. First, the logic of a validation approach based on signal detection theory is described. Then, we demonstrate how signal detection theory can be employed to establish cut-off scores. Finally, the methodology for determining the utility of the checklist evaluation instrument is established.

Validation Concept

The Theory of Signal Detection (TSD) represents a way of characterizing the sensitivity of people making decisions in the face of uncertainty. It has been applied in a wide variety of situations ranging from sensory detection to clinical studies (Hutchinson, 1976). Most prior studies have been concerned with the decision making of individuals rather than the decision making of a system. The present approach differs from most prior studies because it is concerned with validating the decision making of the checklist evaluative system. Specifically, the concept, which is subsequently described, applies TSD to validate a criterion referenced checklist developed to classify graduates of a USAF technical training school.

It is not always clear from the summated score on the checklist whether or not the candidate completing his schooling should be classified as a "master" or a "non-master." Alternatively stated, some candidates should be regarded as signal ("masters") and some as noise ("non-masters"). Unfortunately, the score developed on a candidate by summing the passed items of the checklist is not an unambiguous indicator. In the language of TSD, we have an observation x and must conclude whether it has been drawn from the distribution for signal ("master") against a background of noise ("non-master") or the distribution for noise alone. These distributions are distributions of conditional probabilities and are referred

to as P(x|n), the noise distribution, and P(x|s), the signal distribution. In Figure 7.1, the observation x = 14 is more likely to have come from the noise distribution than from the signal distribution. However, the observation x = 18 is more likely to have come from the signal distribuiton.

Calculations

From Figure 7.2, the hit and false alarm rates (probabilities) may be calculated:

$$P(y|s) = \frac{f(y|s)}{f(s)}$$

$$P(y|n) = \frac{f(y|n)}{f(n)}$$

P (y|s) is the conditional probability of a "hit," i.e., correctly predicting that a student will succeed when placed on the job, and P (y|n) is the conditional probability of a "false alarm." i.e., incorrectly predicting that a student will succeed when placed on the job. It is noted that the predictions of success are made at the end of schooling.

The primary statistic influencing the acceptability of a prediction instrument is its validity. Within the theory of signal detection, as applied here, the equivalent is detection sensitivity (d'). Detection sensitivity is an index of the ability of the system to separate potentially successful from potentially unsuccessful workers. In TSD the resulting distributions are called signal and noise.

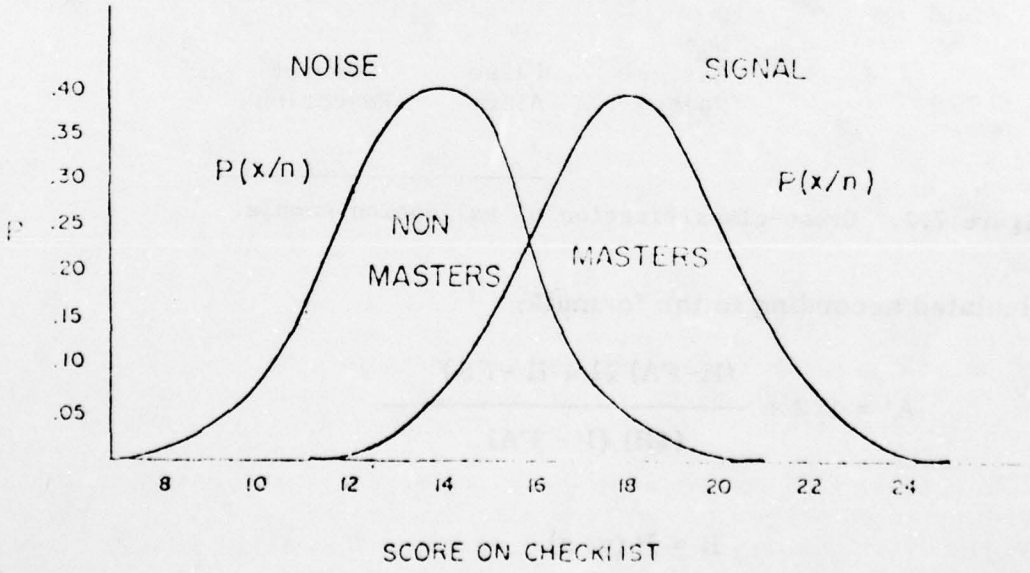


Figure 7.1. Distributions of checklist scores for masters and non-masters.

Dection sensitivity is defined here as the mean of the signal distribution (Mean summated score for "masters") minus the mean of the noise distribution (mean summated score for "non masters") divided by the standard deviation of the noise distribution:

$$d = \frac{M \text{ "masters"} - M \text{ "non masters"}}{\sigma \text{ "non masters"}}$$

Detection sensitivity may be taken directly from the tables prepared by Patricia Elliott (Swets, 1964) by entering with the hit and false alarm probabilities, P (y |n) respectively. Normally, this procedure is employed when scaled numerical data are unavailable. Should the assumptions associated with the calculation of d' be violated, then a nonparametric equivalent (A') of d' may be employed (Pollack, 1970; Pollack & Norman, 1964).

We suggest that each of a set of candidates be classified with the aid of a checklist at the end of his schooling as a "master" or as a "non-master," using an a priori cutting score. After a period of months on the job, there is confirmation or rejection of the prior classification. The four types of outcome are classified in the matrix shown as Figure 7.2.

		Classification at School (Yes) Master	Classification at School (No) Non-Master
Classification in Field	Master (Signal)	Hit	Miss
	Non-Master (Noise)	False Alarm	Correct Rejection

Figure 7.2. Cross-classification of validation sample.

A' is calculated according to the formula:

$$A' = 1/2 + \frac{(H-FA)\,(1 + H - FA)}{(4H)\,(1 - FA)}$$

where:

$$H = P\,(y_i\, s)$$

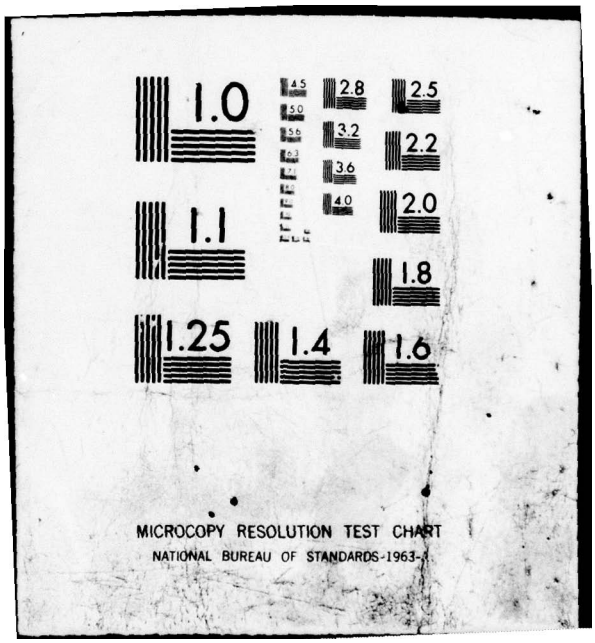
$$FA = P\,(y_i\, n)$$

AD-A074 539 APPLIED PSYCHOLOGICAL SERVICES INC WAYNE PA F/G 5/9
CRITERION REFERENCED TESTING: REVIEW, EVALUATION, AND EXTENSION--ETC(U)
AUG 79 A I SIEGEL, L L MUSETTI, P J FEDERMAN F33615-77-C-0046

UNCLASSIFIED AFHRL-TR-78-71 NL

2 OF 2
AD A074539

END
DATE FILMED
11-79
DDC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

According to Snodgrass (1972), the A' measure is a desirable statistic to compute in situations like the present in which a yes-no experiment with a single payoff matrix and presentation probability are involved.

Likelihood Ratio (Lx)

The likelihood ratio of observation, x, is defined as the ordinate of the signal distribution at x divided by the ordinate of the noise distribution. L(x) values may be obtained by locating the criterion in the signal distribution by the normal z deviate score, zs. This procedure is normally employed when a single set of hit and false alarm rates is available:

$$L(x) = \frac{fs\,(x)}{fn\,(x)} = \frac{P\,(x \mid s)}{P\,(x \mid n)}$$

Interpretation of Results

In the yes-no procedure, a rater is required to divide his continuum of observations into only two parts: He says "Yes" ("master") or he says "No" ("non-master"). His resulting detection sensitivity (d') or the nonparametric equivalent (A') is an index of the validity of the checklist system. The higher the d', the higher the validity of the evaluative system. Low variability among the raters' individual likelihood ratios is an indication of acceptable reliability of the checklist evaluative system. High variability in the likelihood ratio will indicate that the raters are using different criteria for saying "Yes" and "No." High variability in L(x) suggests the need for more precisely establishing the cut-off score for an observer using the checklist. Establishing such scores requires numerical data (e.g., summated scores from the checklists employed and a manipulation of the rater's criterion by using a variety of payoff matrices to force a variety of decisions ranging from liberal to conservative. While it is also possible to manipulate the observer's criterion by varying signal rate (percentage of graduates expected to succeed on the job), this procedure may appear somewhat less realistic to experienced observers.

Establishing Cutoff Scores

The validation concept presented in the previous section will produce a set of summated checklist scores. From this population, a distribution of "masters" (signal) and "non-master" (noise) may be created. Assume that each distribution may be described as follows: Ms = 18), Mn = 14, $\sigma s = \sigma n - 2$, P(s) = 0.5 and P(n) = 0.5 (see Figure 7.1). These data can now form the basis for establishing cutoff scores. A group of observers (about 10) would classify each person rated as a "master" or as a "non-master."

Costs and benefits are associated with right and wrong decisions about graduates who are later placed on the job. An example value system is illustrated in Figure 7.3. Since there are two ways of being correct and two ways of being incorrect, the payoff matrix tells the person examining the summated score that hits and correct rejections have equal value. Similarly, both types of errors are to be weighed equally as costs rather than rewards. (However, symmentry in the payoff matrix is not a necessary condition.)

		Classification at School	
		Master	Non-Master
Classification in Field	Master	+1	-1
	Non-Master	-1	+1

Figure 7.3. Symmetrical payoff matrix.

Each trial may be considered as the examination of a candidate's behavioral checklist by a rater. The subjects are the raters who employ the checklist for a decision making. The stimuli are the checklists, each of which has a summated score. In sensory experiments, highly trained raters are most likely to produce stable results. Here experienced instructors are assumed to be the equivalent of highly trained raters. The raters are told the success probability (e.g., the percentage of graduates found successful on the job = 50%) and are given a payoff matrix to help fix the criterion for the set of trials. Five payoff matrices designed to vary the raters' criterion from conservative (Matrix A) to liberal (Matrix E) are presented (Figure 7.4). The Figure 7.4 matrices were taken from Snodgrass (1972). All raters are expected to make a "Yes" or "No" decision about each of the graduates under five different payoff conditions. Accordingly, if there are 10 raters, 100 persons rated, and 5 payoff matrices, each rater would complete 500 trials and a total of 5,000 data points would be on hand. Table 7.1 presents the sequence with which each rater might be exposed to each payoff matrix. Within each payoff condition the sequence of exposure to the 100 checklists is randomized.

Table 7.1

Sequence with which Raters are Exposed to the Payoff Matrices

	-- Matrix --				
Rater	A	B	C	D	E
1	1	2	3	4	5
2	5	1	2	3	4
3	4	5	1	2	3
4	3	4	5	1	2
5	2	3	4	5	1
6	5	4	3	2	1
7	1	5	4	3	2
8	2	1	5	4	3
9	3	2	1	5	4
10	4	3	2	1	5

A

	Yes	No
S	1	-1
N	-9	9

B

	Yes	No
S	1	-1
N	-2	2

C

	Yes	No
S	1	-1
N	-1	1

D

	Yes	No
S	2	-2
N	-1	1

E

	Yes	No
S	9	-9
N	-1	1

Figure 7.4. Five payoff matrices designed to vary the observers criterion. (Taken from Snodgrass, 1972.) (Reprinted by permission of copyright owner: Life Science Associates, One Fenimore Road, P.O. Box 500, Bayport, New York 11705.)

Table 7.2, taken from Green and Swets (1966), illustrates the kinds of summary statistics calculated for each of rater. P(y|n) constitute the hit and false alarm rates; d' is the detection sensitivity; β is the likelihood ratio and β opt is the value for optimizing correct personnel decisions.

Table 7.2

Data Obtained from a Single Observer in Five Sessions in which Payoffs were Varied (Data from Green & Swets, 1966) Reprinted by permission of copyright owner: John Wiley & Sons, Inc., 605 Third Avenue, New York, New York 10016.

Payoff Matrix	P(Yes/S)	P(Yes/N)	z_S	z_N	$d' = z_N - z_S$	f_S	f_N	$\beta = f_S/f_N$	β_{opt}
A	.245	.040	.690	1.750	1.06	.3145	.0862	3.65	9.00
B	.300	.130	.524	1.126	.60	.3478	.2116	1.64	2.00
C	.695	.335	-.509	.425	.93	.3504	.3644	.96	1.00
D	.780	.535	-.772	.087	.86	.2962	.3974	.75	0.50
E	.975	.935	-1.960	-1.514	.45	.0585	.1268	.46	0.11

Hit and False Alarm Rates

From the hit and false alarm rates, a receiver operating characteristic (ROC) curve may be developed for each observer. Movement along the ROC is criterion change. The point to the lower left in Figure 7.5 represents an instructor with a very conservative criterion, i.e., one who desires a low false alarm rate and therefore sets a high cutoff score for passing a student. As the instructor relaxes his criterion with relationship to the cutoff score, the points move along from ROC from lower left to upper right.

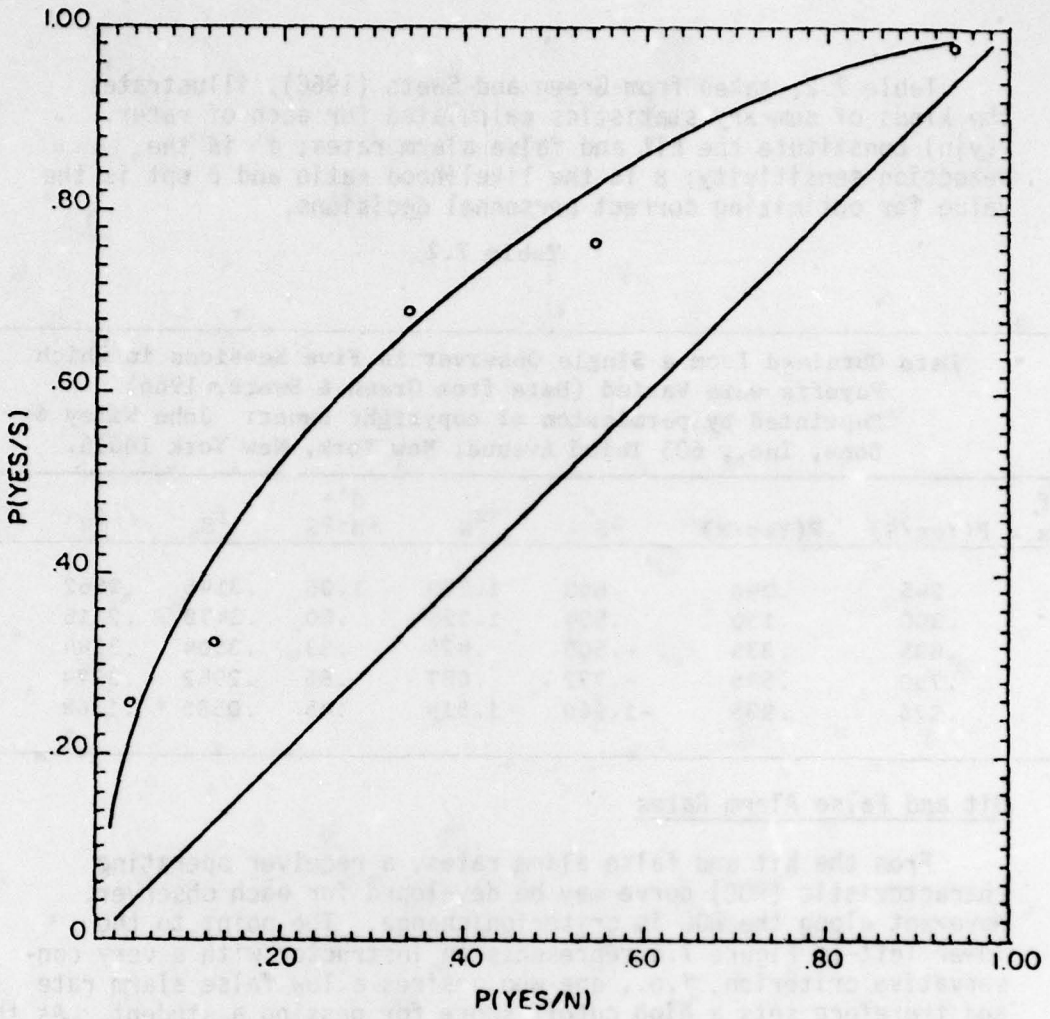


Figure 7.5. ROC curve obtained by varying pafoffs rather than presentation probabilities. The theoretical curve is d'=0.85 and equal variance distributions (from Green & Swets, 1966, p.89). (Reprinted by permission of copyright owner: John Wiley & Sons, Inc., 605 Third Avenue, New York, New York 10016.)

Optimum Value of Likelihood Rater (β Opt)

The value for optimizing correct personnel decisions is calculated by:

$$\beta \text{ opt} = \left[\frac{P(n)}{P(s)}\right]\left[\frac{\text{Value (correct rejection)} + \text{Cost (false alarm)}}{\text{Value (Hit)} + \text{Cost (Miss)}}\right]$$

It is important to remember that the basis for the decision is not the score on the checklist but a transformation of it to a new decision axis, the likelihood ratio. β opt defines the threshold value for the response "Yes". A rater should say "Yes" if L(x) is at least as high as βopt; otherwise he should say "No." An example of the calculation of the optimum β and its corresponding criterion value, for a signal distribution with Ms = 18, a noise distribution with Mn = 14, σs = σn = 2, for various combinations of payoffs and presentation probabilities is presented as Table 7.3, which was taken from Snodgrass (1972).

Table 7.3

Optimum β for Conditions Given in Text (from Snodgrass, 1972)
Reprinted by permission of copyright owner: Life Science Associates, One Fenimore Road, P.O. Box 500, Bayport, New York 11705.

Payoff Matrix (Table 4)	P(S)	βopt	criterion value
A	0.5	$\frac{0.5}{0.5} \times \frac{9 + 3}{1 + 1} = 9$	18.2
B	0.5	$\frac{0.5}{0.5} \times \frac{2 + 2}{1 + 1} = 2$	16.7
C	0.5	$\frac{0.5}{0.5} \times \frac{1 + 1}{1 + 1} = 1$	16.0
D	0.5	$\frac{0.5}{0.5} \times \frac{1 + 1}{2 + 2} = 1/2$	15.3
E	0.5	$\frac{0.5}{0.5} \times \frac{1 + 1}{9 + 9} = 1/9$	13.8

For the calculations illustrated, the criterion for classifying a graduate as a "master" or a "non-master" varies between 13.8 and 18.2. The cutoff chosen depends upon the payoff matrix selected. Standards are expected to vary from time to time. With a symmetric payoff matrix and a market condition where 50% of the candidates are expected to succeed on the job, the cutoff score designed to optimize personnel decisions is 16 items passed.

Reliability and Rater Error

Reliability of the decision system may be calculated by an analysis of the variance of the likelihood ratios (β). This is illustrated in Table 7.4. The variance analytic approach to reliability (Winer, 1962) produces a coefficient interpretable as the most likely correlation between the average of the likelihood ratios produced by the present set of raters and those produced by another set of raters exposed to the same stimuli.

Table 7.4

Likelihood Ratios of 10 Observers Employing Five Different Payoff Matrices

	Observer									
Matrix	1	2	3	4	5	6	7	8	9	10
A										
B										
C										
D										
E										

Note -- To meet the scaling assumptions of ANOVA, β is transformed to β' by the equation $\beta' = 1 - (1/\beta + 1)$.

Rater error tendency may be estimated by calculating the correlation of differences between β' and β' opt for the 10 raters under each payoff condition, as illustrated in Table 7.5. Since β' opt represents the ideal observer and β' represents the actual observer, the resulting coefficient developed from averaging the $\frac{n(n-1)}{2}$ coefficients is an estimate of error for the group.

Table 7.5

Differences between β' and β' opt for 10 Raters Operating under Five Different Payoff Conditions

	Payoff Matrix				
Rater	A	B	C	D	E
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					

Note: In order to meet the assumptions of r, likelihood ratios are transformed accordingly to the function:

$$\beta' = 1 - \frac{1}{\beta + 1}$$

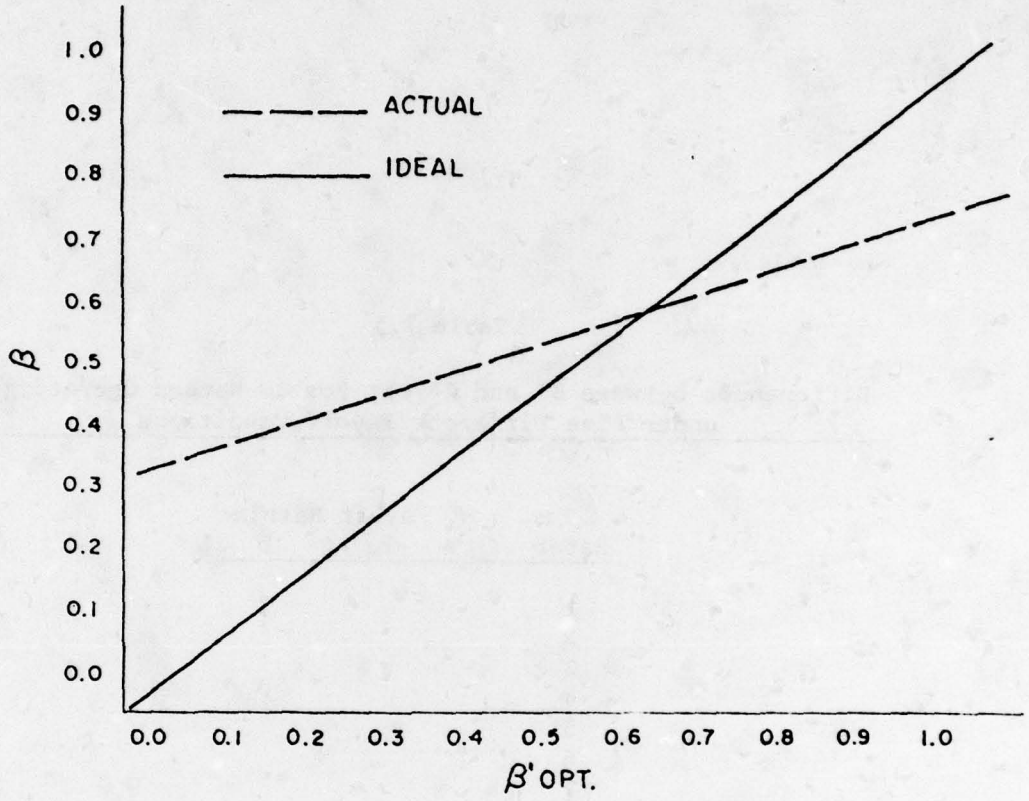


Figure 7.6. Relationship between obtained and optimal decision criteria.

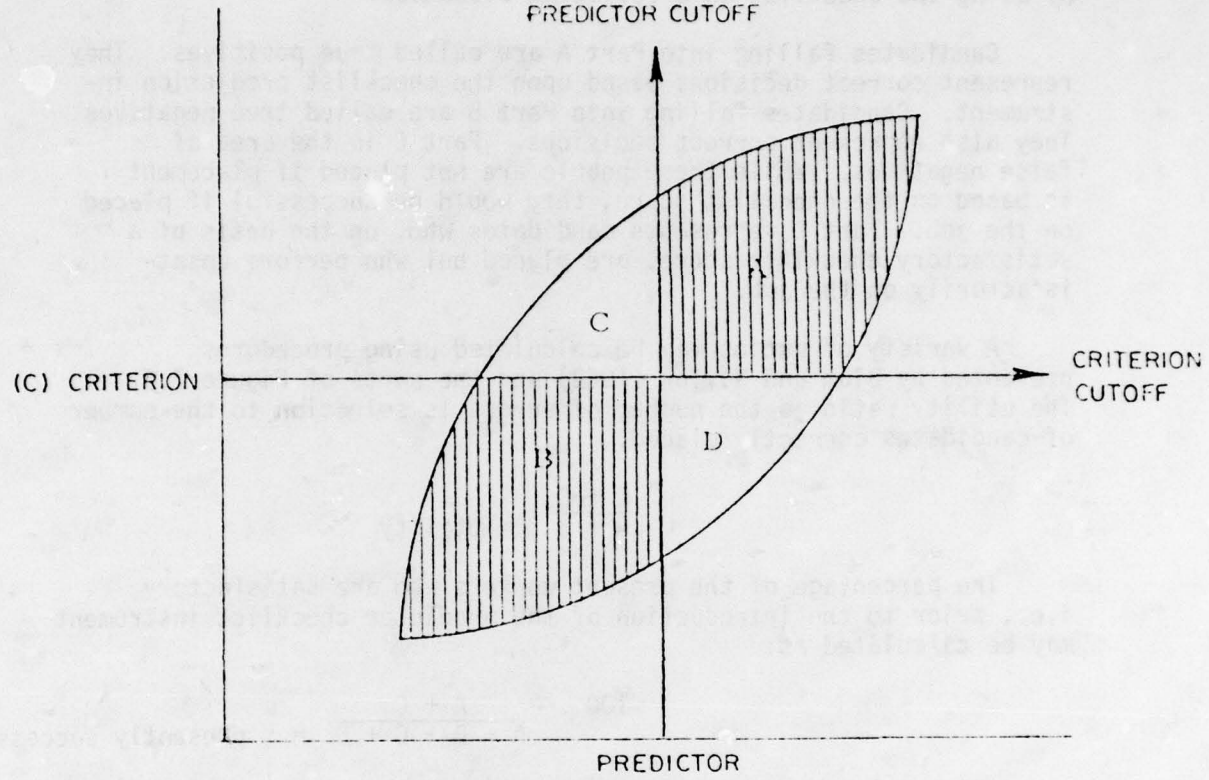


Figure 7.7. Cross classification of candidates.

The relationship between the obtained and the optimal likelihood ratios is derived next. Figure 7.6 illustrates a possible result. The raters overestimate the true value of β at the low end of the scale and underestimate the true value of β at the high end of the scale.

Utility of the Checklist

Figure 7.7 illustrates how the candidates might be classified by using the checklist in a following situation.

Candidates falling into Part A are called true positives. They represent correct decisions based upon the checklist prediction instrument. Candidates falling into Part B are called true negatives. They also represent correct decisions. Part C is the area of false negatives. While these people are not placed if placement is based on the checklist score, they would be successful if placed on the job. Part D represents candidates who, on the basis of a satisfactory checklist score, are placed but who perform unsatisfactorily on the job.

A variety of ratios may be calculated using procedures presented by Blum and Naylor (1968) and the parts of Figure 7.7. The utility ratio is the number of errors is selection to the number of candidates correctly placed.

$$\frac{C + D}{A + B} = \text{Utility}$$

The percentage of the present workers who are satisfactory, i.e., prior to the introduction of the predictor checklist instrument may be calculated as:

$$100 \cdot \frac{A + C}{A + B + C + D} = \% \text{ presently successful}$$

who will be successful if one used the checklist as an aid to selection along with the methods currently being employed may be calculated as:

$$100 \cdot \frac{A}{A + D} = \% \text{ successful with prediction}$$

To the extent that the results of this calculation are greater than those of the prior calculation, the prediction is adding something to the selection process.

Advantages of Suggested Approach

The suggested approach to criterion referenced test verification seems to possess a number of advantages:

- It is less subject to artificial constriction because of restriction of range than the usual correlation statistics.

- It separates the rater's sensitivity to student ability to succeed or fail on the job from the placement of his decision threshold. Accordingly, the approach is both diagnostic and prescriptive. It tells where the observer has erred and allows a basis for corrective action.

- It allows reliability determination in terms of the raters' decision criteria--a proper matter for interrater reliability--in a manner which is compatible with the go-no go philosophy of criterion referenced testing and which is suitably quantified.

- It emphasizes the decision system rather than individual checklist items. The decision system seems to be the subject of interest in the present context.

- It provides a bases for rater training.

- The d' statistic is easily interpretable because it is on a standardized scale.

VIII. PROGRAMMATIC RESEARCH INTO PERFORMANCE CHECKLISTS IN THE USAF

The present chapter summarizes a general program of additional required research into the development and use of performance checklists in the USAF. The suggested research program aims to provide a basis for the proper construction and administration of performance checklists in USAF technical training schools.

The development and use of any test requires adherence to sound test construction and administration procedures. Psychometrically sound test construction and administration procedures serve to increase test reliability and validity; unsound procedures detract from the reliability and validity of test results. No matter what type of test, the major goal of any testing program should be to maximize the reliability and validity of test results. Accordingly, within this research program the emphasis is on studies which will enhance the development and administration of performance checklists so that the checklist results will be reliable and valid.

Orienting the research program in this direction seems warranted on the basis of the results reported in Chapters V and VI of this report. Overall, these studies indicated the possible lack of reliability and validity in performance checklist scores, and suggested that the core of the problem may fall into three areas:

- problems inherent in the method of developing the checklists
- problems inherent in the raters
- problems inherent in the testing conditions

Several test and measurement text books (e.g., Guion, 1965; Cronbach, 1970; Anastasi, 1976) discuss the direct and substantial impact of test content, raters, and test conditions on reliability and validity; hence, these three problem areas are by no means unique to the Air Force. However, they seem to have been studied little in the Air Force performance check context.

Checklist Development

The literature review indicates that prior research has examined the psychometry of (a) item development and writing, (b) item analysis and selection, and (c) item and test scoring format in regard to CR testing. The emphasis of the research was on the development and selection of test items whose content represents the subject matter under test and which will produce test item scores that reference the performance requirements. Studies are needed which will identify the most useful methods to use when constructing such checklists and which will specify the content to be included in such lists.

Item Development

The first step in the development of a training course examination, and of the training course itself, consists of a thorough analysis and understanding of the subject matter to be taught. In the Air Fource, this job analysis is an integral aspect of the Instructional System Development process. However, this analysis takes place at the task level. For performance checklists, specificity is required at the subtask level; each behavioral element of the task should be described. A subtask is the smallest behavioral unit into which a task can be divided without analyzing the separate motions and movements involved. For example, the job analysis of a typing task might produce this subtask statement: "sets the left margins to 3/4" ± 10 before inserting paper." Study is required into methods for extracting subtasks from the task oriented occupational data presently provided by Air Force occupational analytic descriptions. Methods also seem to be needed for identifying critical subtasks within tasks. The end result of such studies would be a set of methods for identifying behaviorally based, objective subtask items. Such items will, presumably, decrease rater errors and produce higher quality ratings than ambiguously written items and items based on abstract psychological traits, (Kavanaugh, 1971; Bernardin, 1977). Such a study appears warranted given the present state-of-the-art of performance checklists. The performance checklists studied by Applied Psychological Services often possessed vague items. For example, in Course A (which contained some of the more detailed performance checklists here reviewed), some items read as follows:

Were accessories inspected properly?

Were adequate safety steps taken during all station operations?

Were realistic priorities developed?

Uses correct procedures during scripted emergency situations?

The underlined words represent ambiguous success criteria.

Within this context, the critical incident approach might constitute one analytic approach of interest. A critical incident approach to subtask analysis might not only provide behaviorally oriented subtask descriptors but also indicate those behaviors that should be avoided. Once the analysis work is completed, a performance checklist might then be composed of items which represent both good and poor subtask performance.

Other methods for deriving the subtask list should also be investigated. The critical incident approach may not provide a complete list of subtask statements. The results of other subtask analytic approaches should be compared with those yielded by a critical incident analysis. The ultimate goal of this set of studies is to develop a preferred method for establishing the subtasks to be included in a performance checklist.

Test Objectives and Length

Two other important considerations in performance checklist development are:

(a) when to test

(b) how many items to include

The major benefits of performance testing are that it allows an absolute evaluation both of training course quality and student skill performance. CR performance test development and administration, however, can be costly in terms of test materials, tools and equipment, and possible remedial training time. With such benefits and costs, a detailed analysis seems required to provide data on the extent to which Air Force schools should use performance tests.

The questions directly facing USAF schools are whether to use performance checklists: (a) for all specialties and all phases within a course, (b) for some specialties but not for others, (c) for some types of output but not for others, (d) for advanced courses but not for basic courses, or (e) not at all. The answers should be derived by the extent that added benefit over cost acrues from this type of testing. Such cost/benefit analyses seem required for performance testing relative to a broad set of Air Force technical training programs.

Once the decision is made to use a performance test, the number of subtask items to include in the performance checklist is another consideration. Either all the subtasks making up the test could be rated or only some subset of subtasks performed during the test could be rated. The methods presented in Chapter IV under the topic of test length could be employed to derive the optimum test length--the binomial and Bayesian models seem useful. However, these have not been tried in the Air Force context. A study of the utility of these methods in the Air Force situation seems warranted.

Item and Test Scoring

The question of how to score the performance checklist in the Air Force remains open. Traditionally, the checklist is scored by placing a checkmark next to each correctly performed subtask item. This method is assumed to reduce rating errors and produce more reliable

and valid ratings. The checklist, however, can be scored through summated ratings or equal-appearing intervals, (Guion, 1965). The effect on reliability and validity of these three checklist formats for performance testing in the Air Force should be examined. Moreover, we note that the empirical work cited earlier in this report found the detailed checklist scoring procedures to be less reliable than an overall rating. Further study of this finding in other specialties seems indicated.

Summated rating scales could be investigated for use along the dimension of the quality of task performance. Alternatively, three, five, or seven scale intervals could be investigated where the last point signifies that the subtask behavior was not performed. Results from application of the various type of scale should be compared to determine whether or not the results are invariant across method of scaling. In other fields, observer judgments have been found to be sensitive to the method of scaling. Such a finding would possess important implications for the interpretation of checklist scores.

Discussion of Checklist Development Studies

The major dependent variables of the suggested research on test construction methods are reliability and validity. It is evident that item development, selection, and scoring procedures may affect the reliability and validity of performance checklist scores. The studies outlined above regarding these test construction variables should provide needed data and procedures for developing more reliable and valid performance checklists in the Air Force.

Rater Characteristics

The effect of the rater on the scoring of a performance checklist is probably a major contributor to the presence or absence of reliability and validity. The literature has generally considered three characteristics of the rater as producting a significant impact on ratings:

- rater qualification
- rater experience
- rater point of view

In the Air Force context, the effect of these three rater characteristics on performance checklist ratings should be investigated.

Rater Qualification

Guion (1965) distinguished rater qualification on the basis of three variables: (1) observation, (2) attitude, and (3) training. In the performance testing context, the raters presumably observe firsthand the testee's performance; as such, all potential raters meet this most important first qualification requirement. Through observation of all test behaviors, no significant effect on performance checklist ratings should occur due to observation alone.

Rater attitude, however, could confound performance checklist ratings. If the rater does not accept the purpose of the performance tests or, for example, if the test is believed to be a waste of time, ratings could be given carelessly. Organizational climate also enters here; instructors may not assign failing ratings because they believe that their supervisors will perceive such ratings as a poor reflection on the instruction. On the other hand, student pressure for good grades may cause the instructor to award inflated ratings. For accurate performance checklist ratings, rater/instructor attitudes are an important aspect of qualification. Given the probable biasing effect of instructor attitudes, the Air Force should examine instructor attitudes toward testing, and attempt to derive factors of the job climate which support positive attitudes.

The third variable is rater training. Guion (1965) considered rater training as a qualification requirement because it develops rating skill. He specified that raters need training on the techniques of rating. Raters require training in at least three areas. First, the raters need to understand the subtask items, instructions, and purposes of the performance checklist. It is especially important that raters are familarized with each behavior represented by the respective subtask items. Second, they need to learn a method for observing test performance, and to know what behaviors to watch for and rate. Third, raters require training in the process of making judgments; specifically, training to avoid or reduce rating errors such as leniency, halo, contrast, logical, proximity, and other rater biases.

Accordingly, development of a program for training raters in the use and administration of performance checklists seems indicated. Such a program would include a section on methods for reducing biases. Past research on rater training has focused on rater biases and has shown some success in reducing rating errors, (Borman, 1975; Latham, Wesley, & Purcell, 1975).

A more basic research area (yet, one which seems quite important) might examine both the process of rater observation and rater judgment. Research could have raters vocalize: (a) what they are watching, and (b) why they give poor or good ratings during actual testing situations. Such data would: (1) help achieve a better understanding of the rating process, (2) suggest additional training areas, and (3) provide insight into the intraindividual structure of the rating process.

Rater Experience

The rater's experiences as a job incumbent, as a training instructor, and as an evaluator using the performance checklist probably influence rating accuracy. It may be that rating errors on performance checklists are, in part, interactive with certain types of experience. Such a possible experience effect was suggested by the study results on rater bias in Chapter VI, the rater (Rater 1) with the least experience teaching the course under study, but the greatest overall teaching experience, showed extreme halo biases.

Research on the performance checklist may want to examine the relationship between rater experience and rating errors, and attempt to determine the reasons for any experience effect.

Rater Point of View

Barrett (1966) defined rater point of view as the style of worker behavior the rater believes shows good job performance. A worker's performance style is the manner in which the worker goes about performing the job. Performance style does not refer to the actual job behaviors; that is, what the worker does, but refers to how the worker generally performs the job behaviors. Barrett suggested that if raters and ratees possess different perceptions on the performance style that constitutes successful performance, then it is likely the ratees will be rated inaccurately.

Adapting Barrett's (1966) point of view to the performance checklist context, it appears possible that rater viewpoint on both the test behaviors themselves and performance style may confound ratings. If viewpoints differ, the likelihood that separate raters will rate the same behaviors differently is high; also ratees may receive low ratings not necessarily because they performed poorly but because their style disagrees with the rater's style.

Research in this area seems warranted. First, it seems important to know whether or not different work styles exist across and within various Air Force technical specialties. Second, if such styles are identified by research, then the effect of the styles on the ratings should be investigated. Such work would include both interrater reliability and rating accuracy investigations. Such data could also be helpful in correcting performance standard misconceptions.

Viewpoint on performance style could also interact with performance checklist item characteristics. Such a result seems logical. For example, one type of rating scale may be more affected by performance style viewpoint than another. Identification of such covariance, if present, might represent another research avenue.

Discussion of Rater Characteristics Research

Besides performance checklist content and format and rater characteristics, the conditions of the performance test and the rating procedure will also likely affect test reliability and validity. Such effects are also quite likely to be highly interactive. It is also possible that the performance style could serve as a moderating variable such that a given rater-checklist combination will be effective for a rater with one performance style but not for a rater with another performance style.

Testing Conditions

Testing, whether in the form of a paper-and-pencil test or apparatus type performance test, requires standardized conditions. As for any other measurement, test results should not be uninterpretable because the results are confounded with the conditions of the test. Nonstandardized and uncontrolled test conditions confound the scores of the best developed test and may cause the most accurate rater to error. Anastasi (1976) emphasized the biasing effect of poor test conditions on test scores. She wrote:

> Even apparently minor aspect of the testing situation may appreciably alter performance. Such a factor as the use of desks or chairs with desk arms, for example, proved to be significant...There is also evidence to show that the type of answer sheet employed may affect test scores (p.33).

With the possible major impact of even a minor factor on test results, research on the effect of different test conditions on performance checklist results seems required. One area of research could estimate the error variance associated with different test conditions, and the ratings corrected for the test conditions could be calculated much in the same way ratings can be corrected for rater bias (see Chapter VI).

In the performance checklist, at least three test condition variables seem important for investigation:

- instructions
- rater behavior
- test materials, tools and equipment

Each of these factors, if uncontrolled and left to vary from test to test, will probably confound a performance checklist's ratings--making them less reliable and valid. For example, rater behavior toward the ratee prior and during the performance test could affect the performance of the individual being tested. The difference of a "warm" versus "cold" or "rigid and aloof" versus a "natural manner" by the rater on tested score should be investigated in the Air Force technical training situation. Different instructions or coaching by the rater; nonstandard tools, worn materials represent conditions which might influence test results.

Research into the effects of each of these variables on performance check results seems indicated along with the development of a set of recommendations for minimizing the effects of such variables on test results.

REFERENCES

Abrams, A.J., & Pickering, E.J. A checklist for use by a trainer-evaluator to assess sonarman proficiency or test equipment. (Tech. Bull. 62-7). San Diego, Calif.: Bureau of Naval Personnel, 1962.

Alkin, M.C. "Criterion-Referenced Measurement" and other such terms. In C.W. Harris, M.C. Alkin, & W.J. Popham (Eds.). Problems in criterion-referenced measurement. Los Angeles: Center for the Study of Evaluation, University of California Graduate School of Education, 1974.

Alvord, D.J., & Buttingham, B.E. Evaluating performance on national assessment objectives: norm-reference and criterion-reference interpretations. Journal of Education Research, 1974, 68, 59-61.

Anastasi, A. An empirical study of the applicability of sequential analysis to item selection. Educational and Psychological Measurement, 1953, 13, 3-13.

Anastasi, A. Psychological testing (4th Ed.). New York: Macmillan, 1976.

Barrett, R.S. Performance Rating. Chicago: Science Research Associates, 1966.

Bernardin, H.J. Behavior expeciation scales versus summated scales: A fair comparison. Journal of Applied Psychology, 1977, 62, 422-427.

Block, J.H. Student learning and the setting of performance standards. Educational Horizons, 1972, 183-191.

Blum, M.L., & Naylor, J.C. Industrial Psychology: Its theoretical and social foundations. New York: Harper and Row, 1968.

Borman, W.C. Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. Journal of Applied Psychology, 1975, 60, 556-560.

Borman, J.H. On the theory of achievement test items. Chicago: The University of Chicago Press, 1970.

Brown, B. Delphi Process: A methodology used for the elicitation of opinions of experts. Santa Monica: Rand Corp., 1968.

Campbell, D.T. & Stanley, J.C. Experimental and quasi-experimental designs for research. Chicago: Rand-McNally, 1963.

Campbell, J.P., Dunnette, M.D., Lawler, E.E., III, & Welck, K.E., Jr. Managerial Behavior, Performance, and Effectiveness. New York: McGraw Hill, 1979.

Chesire, L., Saffir, M. & Thurstone, L.I. Computing diagrams for the tetrachoric correlation coefficient. Chicago: University of Chicago Press, 1938.

Cox, R.C. Evaluative aspects of criterion referenced measures. In W.J. Popham (Ed.), Criterion referenced measurement: An introduction. Englewood Cliffs, New Jersey: Educational Technology Publications, 1971.

Cox, R.C., & Vargas, J.S. A comparison of item selection techniques for norm-referenced and criterion-referenced tests. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, February 1966.

Crehan, K.D. Item analysis for teacher made mastery tests. Journal of Educational Measurement, 1974, 11, 255-262.

Cronbach, L.I. Essentials of psychological testing. New York: Harper & Row, 1979.

Cronbach, L.I., Gleser, G.C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley, 1972.

Dalkey, N. Delphi. Santa Monica: RAnd Corp., 1967.

Dalkey, N. An experimental study of group opinion. Futures, 1969, 2.

Dalkey, N., & Helmer, O. The use of experts for the estimation of bombing requirements: A project Delphi experiment. Santa Monica: Rand Corp., 1962.

Dayton, C.M., & Macready, G.B. A probabilistic model for validation of behavioral hierarchies. Psychometriks, 1976, 41, 189-204.

Department of the Army Scoring Booklet (11B2177, 11B3177, 11B4177), Washington, D.C.: Headquarters, Department of the Army, 1977.

Ebel, R.L. Must all tests be valid? American Psychologist, 1961, 16, 640-647.

Emrick, J.A. An evaluation model for mastery testing. Journal of Educational Measurement, 1971, 8, 321-326.

Epstein, K.I., Steinheiser, F.E., Macready, G.B., & Mirabella, A. Methods and models for criterion-referenced testing. Unpublished manuscript, 1977 (Available from authors at the Army Research Institute for the Behavioral and Social Sciencies, Arlington, VA.).

Epstein, K.I., & Steinheiser, F.H. A Bayesian method for evaluating trainee proficiency. Procedings of the 8th Naval Training Equipment Center/ Industry Conference, Orlando, Florida, November 1975.

Gagne, R.M. Some notes on criterion-referenced measurement. Unpublished manuscript, 1969 (Available from author at Florida State University, Tallahassee, Fla.).

Garvin, A.D. The applicability of criterion-referenced measurement by content area and level. In W.J. Popham (Ed.), Criterion-referenced measurement: An introduction. Englewood Cliffs, N.J.: Educational Technology Publications, 1971.

Glaser, R. Instructional technology and the measurement of learning outcomes: Some questions. American Psychologist, 1963, 18, 519-521.

Glaser, R. A criterion-referenced testing for the measurement of educational outcomes. In R.A. Weisgerber (Ed.), Instructional process and media innovation. Chicago: Rand McNally, 1968.

Glaser, R. & Cox, R.C. Criterion-referenced testing for the measurement of educational outcomes. In R.A. Weisgerber (Ed.), Instruction process and media innovation. Chicago: Rand McNally, 1968.

Glaser R., & Nitko, A.J. Measurement in learning and instruction. In R.L. Thorndike (Ed.), Educational Measurement. Washington, D.C.: American Councial on Education, 1971.

Gorth, W.P., & Hambleton, R.K. Measurement considerations for criterion-referenced testing and special education. The Journal of Special Education, 1972, 6, 303-314.

Green, D.M., & Swets, J.A. Signal detection theory and psychophysics. New York: Wiley, 1966.

Gronlund, N.E. Preparing criterion-referenced tests for classroom instruction. New York: Macmillian, 1973.

Guilford, J.P. Psychometric methods (2nd Ed.), New York: McGraw-Hill, 1954.

Guilford, J.P. Fundamental statistics in psychology and education (4th Ed.), New York: McGraw-Hill, 1965.

Guion, R.M. Personnel testing. New York: McGraw-Hill, 1965.

Guttman, L. A basis for scaling qualitative ideas. American Sociological Review, 1944, 9, 139-150.

Haladyna, T.M. Effects of different samples on item and test characteristics of criterion-referenced test. Journal of Educational Measurement, 1974, 11, 93-99.

Hambleton, R.K., & Gorth, W.P. Criterion-referenced testing: Issues and applications (Technical Report No. 13). Amherst, Massachusetts: Center for Educational Research, University of Massachusetts, September 1971.

Hambleton, R.K., & Novick, M.R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Harris, C.W. An interpretation of Livingston's reliability coefficient for criterion-referenced tests. Journal of Educational Measurement, 1972, 9, 27-29

Harris, M.W., & Steward, D.M. Application of classical strategies to criterion-referenced test construction. Paper presented at the meeting of the American Educational Research Association, New York, 1971.

Helmer, O. Analysis of the future: The Delphi method. Santa Monica: Rand Corp., 1967.

Hsu, T. Empirical data on criterion-referenced tests. Paper presented at the meeting of the American Educational Research Association, New York City, February 1971. (ERIC Document Reproduction Service No. ED 050 139)

Hutchinson, T.P. The usefulness of signal detection theory in the analysis of ordinal data from diverse fields. London: University College, 1976.

Ivens, S.H. An investigation of item analysis, reliability, and validity in relation to criterion-referenced tests. Unpublished doctoral dissertation, Florida State University, 1970.

Kavanaugh, M.J. The content issue in performance appraisal: A review. Personnel Psychology, 1971, 24, 653-669.

Klein, S.P., & Kosecoff, J. Issues and procedures in the development of criterion-referenced tests. Princeton, N.J.: Educational Testing Service, 1973. (ERIC Document Reproduction No. ED 013 371).

Knerr, C.S., & Epstein, K.I. Sequential analysis for individual proficiency decisions. Paper presented at the meeting of the Military Testing Association, Gulf Shores, Alabama, October 1976

Latham, G.P., Wesley, K.N., & Pureell, E.D. Training managers to minimize rating errors in the observation of behavior. Journal of Applied Psychology, 1975, 60,550-555.

Livingston, S.A. Criterion-referenced applications of classical test theory. Journal of Educational Measurement, 1972, 9, 13-26.

Lord, F.M. Some test theory for tailored testing. In W.H. Holtzman (Ed.), Computer-assisted instruction, testing, and guidance. New York; Harper & Row, 1970.

Lunney, G.H. Using analysis of variance with a dichotomous dependent variable: An empirical study. Journal of Educational Measurement, 1979, 7, 263-269.

Martino, J. The precision of the Delphi estimates. Technological Forecasting and Social Change, 1972, 1, 293-299.

McCormick, E.J., & Tiffin, J. Industrial Psychology (6th Ed.). Englewood Cliffs, N.J.: Prentice-Hall, 1974.

Millman, J. Passing scores and test length for domain-referenced measures. Review of Educational Research, 1973, 43, 205-216.

Pollack, I. A nonparametric procedure for evaluation of true and false positive. Behavior Research Methods and Instrumentation, 1970, 2, 155-156.

Pollack, I., & Norman, D.A. A nonparametric analysis of recognition experiments. Psychonomic Science, 1964, 1, 125-126.

Popham, W.J. Indices of adequacy for criterion-referenced test items. In W.J. Popham (Ed.), Criterion-referenced measurement: An introduction. Englewood Cliffs, N.J.: Educational Technology Publications, 1971.

Popham, W.J., & Husek, T.R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

Rahmlow, H.F., Mathews, J.J., & Jung, S.M. An empirical investigation of item analysis in criterion-referenced tests. Paper presented at the meeting of American Educational Research Association - National Council on Measurement in Education, Minneapolis, March 1970.

Richlin, M., Federman, P., & Siegel, A.I. Post training performance criterion development and application: development and application of a TBCL criterion to the SESR program for jet aviation machinist's mates. Wayne, Pa.: Applied Psychological Services, 1958.

Richlin, M., Siegel, A.I., Schultz, D.G., & Benson, S. Post training performance criterion development and application: development and application of a TBCL criterion to the SESR program for aviation electronics technicians. Wayne, Pa.: Applied Psychological Services, 1960.

Roundabush, G.E., & Green, D.R. Aspects of a methodology for creating criterion-referenced tests. Paper presented at the meeting of the National Council for Measurement in Eduction, Chicago, April 1972.

Schultz, D.G., & Siegel, A.I. Generalized Thurstone and Guttman scales for measuring technical skills in job performance. journal of Applied Psychology, 1961, 45, 137-142.

Shavelson, R.J., Block, J.H., & Ravitch, M.M. Criterion-referenced testing comments on reliability. Journal of Educational Measurement, 1972, 9, 133-137.

Shoemaker, D.M. Criterion-referenced measurement revisited. Educational Technology, 1971, X1, 61-62.

Shoemaker, D.M. Improving CR measurement. The Journal of Special Education, 1972, 6, 315-323.

Shriver, E.L., & Foley, J.P., Jr. Evaluating maintenance performance: The development and tryout of criterion-referenced job task performance tests for electronic maintenance (AFHRL-TR-74-57 (11), Part 1). Brooks Air Force Base, Tex.: Headquarters Air Force Human Resources Laboratory, September 1974. (NTIS No. AD-A 004-845).

Siegel, A.I. The checklist as a criterion of proficiency. In W.W. Ronan & E.P. Prien (Eds.), Perspectives on measurement of human performance, New York: Appleton-Century-Crofts, 1971.

Siegel, A.I., Bergman, B.A., & Lambert, J. Nonverbal and culture fair performance prediction procedures: II. Initial validation. Wayne, Pa.: Applied Psychological Services, 1973.

Siegel, A.I., Richlin, M., & Federman, P. Post-training performance criterion development and application: development and application of TBCL criteria to the SESR program for the air controlman and the parachute rigger ratings. Wayne, Pa.: Applied Psychological Services, 1958.

Siegel, A.I., Richlin, M., & Federman, P. A comparative study of "transfer through generalization" and "transfer through identical elements" in technical training. Journal of Applied Psychology, 1960, 44, 27-30.

Siegel, A.I., Schultz, D.G. Post-training performance criterion development and application: A further study into technical performance checklist criteria which meet the Thurstone and Guttman scalability requirements. Wayne, Pa.: Applied Psychological Services, 1960.

Siegel, A.I., Schultz, D.G., Fischl, M.A., & Lanterman, R.S. Absolute scaling of job performance. Journal of Applied Psychology, 1968, 52, 313-318.

Simon, G.B. Comments on "implications of criterion-referenced measurements." Journal of Educational Measurement, 1969, 6, 259-260.

Snodgrass, J.G. Theory and experimentation in signal detection. Baldwin, New York: Life Science Associates, 1972.

SQT - A Guide for leaders (no. 350). Washington, D.C.: Headquarters, Department of the Army, 1977.

Standards for educational and psychological tests. Washington, D.C.: American Psychological Association, 1974.

Swaminathan, H., Hambleton, R.K., & Algina, J. Reliability of criterion-referenced tests: A decision-theoretic formulation. Journal of Eductional Measurement, 1974, 11, 263-267.

Sweezey, R.W., & Pearlstein, R.B. Guidebook for developing criterion-referenced tests. Arlington, Va.: U.S. Army Research Institute for the Behavioral and Social Sciences, 1975. (NTIS No. Ad-A 014987).

Swets, J.A. (Ed.) Signal detection and recognition by human observers. New York: Wiley, 1964.

Taylor, C.H., & Russell, J.T. The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. Journal of Applied Psychology, 1939, 23, 565-578.

Tiffin, J., & Hudson, T.W. Comparison of sequential and convertical item analysis when used with primary groups varying in size and composition. Educational and Psychological Measurement, 1956, 16, 333-344.

Unks, N.J. An investigation of validity and reliability concepts for criterion-referenced measurement. Unpublished master's thesis, University of Pittsburgh, 1971.

U.S. Department of Labor: Manpower Administrator Dictionary of Occupational Titles (3rd Ed.). Washington, D.C.: U.S. Government Printing Office, 1965.

U.S. Department of Defense. Military-Civilian Occupational Source Book. Universal City, Texas: Armed Forces Vocational Testing Group, 1975.

U.S. Air Force. AFM 39-1: Airman classification - Manual (Rev.Ed.). Washington, D.C.: Headquarters U.S. Air Force, 1969.

Wald, A. Sequential analysis. New York: Wiley, 1947.

Waters, B.K. An empirical investigation of the stratified adaptive computerized testing model. Applied Psychological Measurement, 1977, 1, 141-152.

Winer, B.J. Statistical principles in experimental design. New York: McGraw-Hill, 1962.

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

☆U.S. GOVERNMENT PRINTING OFFICE: 1979- 671-056/97