Technical Report 416

(12)

ADA082751

CONTEMPORARY VIEWS ON CRITERION-REFERENCED TESTING

R. W. Swezey, R. B. Pearlstein, W. H. Ton
Applied Science Associates

and

Angelo Mirabella
Army Research Institute

20000727322

SIMULATION SYSTEMS TECHNICAL AREA

Reproduced From
Best Available Copy

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

October 1979

Approved for public release; distribution unlimited.

80 4 7 174

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

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

JOSEPH ZEIDNER
Technical Director

WILLIAM L. HAUSER
Colonel, U S Army
Commander

Research accomplished for the
Department of the Army

Applied Science Associates

NOTICES

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

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

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

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: Technical Report 416

2. GOVT ACCESSION NO.:

3. RECIPIENT'S CATALOG NUMBER: ARI TR-416

4. TITLE (and Subtitle): CONTEMPORARY VIEWS ON CRITERION-REFERENCED TESTING

5. TYPE OF REPORT & PERIOD COVERED: --

6. PERFORMING ORG. REPORT NUMBER: --

7. AUTHOR(s): Robert W. Swezey, Richard B. Pearlstein, William Ton [redacted] Angelo Mirabella ~~(Army Research Institute)~~

8. CONTRACT OR GRANT NUMBER(s): DAHC 19-74-C-0018

9. PERFORMING ORGANIZATION NAME AND ADDRESS: Applied Science Associates, Inc. Box 158 Valencia, PA 16059

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 2Q164715A757

11. CONTROLLING OFFICE NAME AND ADDRESS: U.S. Army Research Institute for the Behavioral and Social Sciences 5001 Eisenhower Avenue, Alexandria, VA 22333

12. REPORT DATE: Oct[redacted]79

13. NUMBER OF PAGES: 94

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): --

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): --

18. SUPPLEMENTARY NOTES:

19. KEY WORDS (Continue on reverse side if necessary and identify by block number):
Performance tests
Criterion Referenced Tests (CRT)
Test development
CRT construction

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

This review of the technical and theoretical literature in the area of criterion-referenced testing (CRT) considers a number of areas in CRT development and application. Discussed, in turn, are questions of CRT reliability and validity in both practical and theoretical areas. Different methods of CRT construction are reviewed, as is the question of simulation fidelity (e.g., the estimate to which CRTs can and should mirror real-world performance conditions). Discussion is directed to the use of CRTs in mastery

(Continued)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

032 170

Item 20 (Continued)

learning contexts and to test materials development and item sampling. Diagnostic uses of CRTs and the establishment of cut-off scores are considered. Uses of CRTs in public education and military context are reviewed. Finally a position is set forth on general and theoretical aspects of CRT construction and use.

Technical Report 416

CONTEMPORARY VIEWS ON CRITERION-REFERENCED TESTING

R. W. Swezey, R. B. Pearlstein, W. H. Ton
Applied Science Associates

and

Angelo Mirabella
Army Research Institute

Submitted by:
Frank J. Harris, Chief
SIMULATION SYSTEMS TECHNICAL AREA

Approved by:

A. H. Birnbaum, Acting Director
ORGANIZATIONS AND SYSTEMS
RESEARCH LABORATORY

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES
5001 Eisenhower Avenue, Alexandria, Virginia 22333

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

October 1979

Army Project Number
2Q164715A757

Performance Test
Development

Approved for public release; distribution unlimited.

ARI Research Reports and Technical Reports are intended for sponsors of R&D tasks and for other research and military agencies. Any findings ready for implementation at the time of publication are presented in the last part of the Brief. Upon completion of a major phase of the task, formal recommendations for official action normally are conveyed to appropriate military agencies by briefing or Disposition Form.

FOREWORD

This publication is part of a larger program on criterion-referenced, performance-oriented evaluation being conducted by the U.S. Army Research Institute for the Behavorial & Social Sciences (ARI). A major goal of the program has been to develop procedures for applying CRT Theory to a variety of training situations, including crew and tactical training.

This report summarizes an analysis of the state-of-the-art on criterion-referenced testing, which preceded the preparation of a test construction handbook ("Guidebook for Developing Criterion-Referenced Tests. ARI Report, P-75 1, August 1975). Related efforts have included scoring procedures for performance-based training in tank gunnery (IDOC) and experiments to compare the accuracy of several CRT models in fitting empirical data (METTEST).

ARI research in this area is conducted as an in-house effort augmented by contracts with organizations selected as having unique capabilities and facilities for research in a specific area. The present study was conducted by personnel of the Army Research Institute and Applied Sciences Associates, Inc., under Contract Number DAHC-19-74-C-0018, and was responsive to the requirements of RDTE Project 2Q164715A757, Training Systems Applications.

JOSEPH ZEIDNER
Technical Director

BRIEF

Requirement:

To analyze current state-of-the-art in criterion-referenced testing, and to establish positions on various test contruction issues.

Procedure:

A review and analysis of the literature related to criterion-referenced testing was undertaken. This review included military field manuals, technical reports and personal communications, as well as professional journals. Major topics included definition and use of CRTs, reliability/validity, and test construction.

Findings:

The findings consisted of a set of position statements under four major headings:

1. Design considerations and CRT use
2. Construction methodology and related issues
3. CRT administration and scoring
4. Reliability and validity

Utilization of Findings:

The findings of this study provided a major basis for the preparation of ARI Report P 75 1, "Guidebook for Developing Criterion-Referenced Tests."

Abstract

A review of the technical and theoretical literature in the area of Criterion-Referenced Testing (CR Testing) is presented. A number of areas in CRT development and application are considered. Discussed, in turn, are questions of CRT reliability and validity in both practical and theoretical areas. Different methods of CRT construction are reviewed, as is the question of simulation fidelity (e.g., the estimate to which CRTs can and should mirror real-world performance conditions). Discussion is directed to the use of CRTs in mastery learning contexts and to test materials development and item sampling. Diagnostic uses of CRTs and the establishment of cut-off scores are considered. Uses of CRTs in public education and military context are reviewed. Finally a position is set forth on general and theoretical aspects of CRT construction and use.

Introduction

The distinction between norm-referenced measurement (NRM) and criterion-referenced measurement (CRM) has been aptly illustrated by Popham and Husek (1969) using the analogy of a dog owner who wants to keep his dog in the back yard. The owner finds out how high the dog can jump (a criterion-referenced test) and builds a fence high enough to keep the dog in the back yard. How high the dog can jump compared to other dogs (a norm-referenced test) is irrelevant.

Beginning with Glaser (1963), a number of researchers have made similar distinctions. Folley (1967) for example, has discussed this distinction in the areas of predictive testing and achievement testing. In the case of predictive testing the standard is relative; the results attempt to show how any single individual compares with all other individuals who have taken the test. In achievement testing however, the standard is absolute. The results attempt to show the extent to which an individual has learned a specific set of behaviors. Discrimination among individuals is of secondary importance.

Glaser and Nitko (1971, p. 653) have defined a criterion-referenced test (CRT) as "one that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards", a definition which has been slightly expanded by Livingston (1972, p. 13): "criterion-referenced [is] used to refer to any test for which a criterion score is specified without reference to the distribution of scores of a group of examinees." Common to all definitions is the notion that a well-defined content domain and the development of procedures for generating appropriate

samples of test items are important. Lyon (1972) argues for the use of CRM as a vital part of training quality control:

> . . . quality control requires absolute rather than relative criteria. Scores and grades must reflect how many course objectives have been mastered rather than how a student compares with other students.

Carver (1974) has classified tests as primarily "psychometric", if they focus on individual differences, or primarily "edumetric", if they are designed for sensitivity to within-individual gains. Psychometrically designed tests, in Carver's view, may not be suitable for measuring individual gains, even though they are often used for that purpose. Carver's classification can be applied to the CRT-NRT distinction: Generally, NRTs are psychometric tests, while CRTs are predominantly edumetric.

For the purposes of this review, a CRT will be defined as a test from which the score of an individual is interpreted against an external standard (e.g., a standard other than the distribution of scores of other examinees). Further, CRTs are tests whose items are operational definitions of behavioral objectives.

The literature of psychology and education contains reference to no more persistent problem than that of criterion specification (Ronan & Prien, 1973). Still, no generally accepted method for selecting relevant, reliable, and practical criterion measures exists today.

Use

The contemporary interest in mastery learning has led to a growing interest in the use of CRM. CRTs can be used to serve at least two purposes:

1. They can be used to provide specific information about the performance levels of individuals on instructional objectives. This information can be used to support a decision about "mastery" of a particular objective (Block, 1971).
2. They can be used to evaluate the effectiveness of instruction. NRTs given at the end of a course are less useful for making evaluative decisions about instructional effectiveness, since they are not derived from particular task objectives. CRTs are, however, useful for this purpose because of the specificity of the results to the task objectives (Lord, 1962; Cronbach, 1963; Shoemaker, 1970, 1970b; Hambleton, Rosinelli and Garth, 1971).

Popham (1973) has pointed out a basic concern with the instrument itself:

> We have not yet made an acceptable effort to delineate the defining dimensions of performance tests, in terms of their content, objectives, post-test nature, background information level, etc. Almost all of the recently developed performance tests have been devised more or less on the basis of experience and instruction.

Ebel (1971) has posed a series of arguments against the use of CRM in education. Ebel points out with some justification that CRTs do not tell us <u>all</u> we need to know about educational achievement, pointing out that they are not efficient at discovering relative strengths and deficiencies. Ebel appears to confuse the concept of mastery of material with the practice of using percentile grades as pass-fail measures, and does not address the

notion that CRTs as currently constructed are the result of the application of a carefully thought out analysis and development system.

Klein and Kosecoff (1973) have provided a useful figure summarizing uses of CRTs. They classified CRT uses as to planning, types of decision, and research; and as to individual, group, or program evaluation purposes.

Reliability and Validity

As Glaser and Nitko (1971) point out, the appropriate technique for an empirical estimation of CRT reliability is unclear. Popham and Husek (1969) suggest the traditional NRT estimates of internal consistency and stability are not often appropriate because of their dependency on total test score variability. CRTs typically are interpreted in an absolute fashion, hence, variability is drastically reduced. This section will critically examine a number of studies which have addressed the question of reliability. The question of validity is inextricably mingled with the reliability issue and also presents many facets of opinion and theory. Various positions concerning reliability and validity will be discussed in turn.

Reliability

Smith (1965) has proposed that reliability of test results be assessed by the range of variation of test results:

$$\pm\ 1.96\sqrt{\frac{pq}{N}}$$

where: p = proportion passing,

q = proportion failing

N = number taking test.

Smith suggested that this statistic specifies the range within which 95%

of classes would fall by chance variation. Unfortunately, Smith's range of variation statistic is of limited value, since, for it to be statistically valid, all classes would have to consist of the same number of students. Further, as Smith noted, the range only holds when ". . . the next class is of comparable aptitude, and. . . no significant change, for better or for worse, has been made in the instruction. . ."

Cox and Vargas (1966) compared the results obtained from two item analysis procedures using both pre-test and post-test scores; a Difference Index (DI) was obtained in two ways. A post-test minus pre-test DI was obtained by subtracting the percentage of students who passed an item on the pre-test from the percentage who passed that item on the post-test. A DI was also obtained for each item in the more conventional manner: The distribution of scores on the post-test was divided into thirds and the percentage of students in the lower third on the overall test who passed the item was subtracted from the percentage of students in the upper third who passed the same item. The Spearman Rhos obtained between the two DIs were of a moderate order. The authors concluded that their DI differed sufficiently from the traditional method to warrant its use with CRTs.

Hambleton and Garth (1971) replicated the work of Cox and Vargas (1966) and found that the choice of statistic does indeed have a significant effect on the selection of test items. The change in item difficulty from pre- to post-test seems particularly attractive where two test administrations are possible. Unfortunately, this method uses statistical procedures dependent on score variability which are questionable for CRM (Popham and Husek, 1969; Randall, 1972), particularly if the method is to be employed for item selection

(Oakland, 1972).

Livingston (1972a) acknowledges Popham and Husek's (1969) comment that "the typical indexes of internal consistency are not appropriate for criterion-referenced tests". Nevertheless, Livingston (1971, 1972a, 1972c) has suggested that the classical theory of true and error scores can be used to determine CRT reliability. Livingston (1972a, 1972c) points out that "when we use criterion-referenced measures we want to know how far. . .[a] score deviates from a fixed standard." In Livingston's model, ". . . each concept based on deviations from a mean score. . . [is] replaced by a corresponding concept based on deviations from the criterion score." In this view, ". . . criterion-referenced reliability can be interpreted as a ratio of mean squared deviations from the criterion score."

Livingston cites Lord and Novick's (1968) definition of norm-referenced test reliability as the squared correlation between observed score and true score. Based on this definition, Livingston defines criterion-referenced reliability as "the squared criterion-referenced correlation between observed score and true score". Using algebraic proofs, Livingston demonstrates that this criterion-referenced correlation equals the ratio of mean squared deviations of true scores from the criterion score to the mean squared deviation of observed scores from the criterion score. This ratio is, of course, predicated on the assumption that one can substitute for variance (a concept based on differences from the mean) by using mean squared deviation of scores from the criterion score. If this view is accepted, a number of useful relationships are provided; for instance, the further a mean score is from the criterion score, the greater the criterion-referenced reliability of the test

for that particular group. In effect, moving the mean score away from the criterion score has the same effect on criterion-referenced reliability that increasing the variance of true scores has on norm-referenced reliability. In other words, errors of misclassification of the false positive variety can be minimized by accepting as true masters the group that comfortably exceeds the required criterion level.

According to Livingston, ". . . the farther any person's obtained score is from the criterion score, the more confident we can be in saying that his true score is on the same side of the criterion score". That is, errors of measurement will less likely cause misclassification when the observed score is comfortably distant from the criterion score.

Another point is that if we accept Livingston's model, then the criterion-referenced correlation between two tests depends on the difficulty level of the tests for the particular group involved. Two tests can have a high correlation only if each is of similar difficulty for a group of examinees. This limits the computation of inter-item correlations, as it is often difficult to ensure equal difficulty levels. Livingston's (1971, 1972c) paper included formulas for criterion-referenced applications of the Spearman-Brown formula, coefficient Alpha, and correction for attenuation, as well as the derivation of his basic reliability formula.

Regarding Livingston's (1972a) proposal that the psychometric theory of true and error scores could be adapted to CRM, Oakland (1972) commented that the procedures seemed viable but that the conditions under which they could be used--i.e., availability of suitable NRT measures of criterion behaviors, and multi-item rather than single item CRTs--were overly restrictive.

Harris (1972) objected to Livingston's (1972a) application of classical psychometric theory to CRT, pointing out that whether Livingston's coefficient or a traditional one is applied, the standard error of measurement remains the same. The fact that Livingston's coefficient is usually the larger does not mean a more dependable determination of whether or not a true score falls above or below the criterion score. As a rebuttal, Livingston (1972b) indicated that Harris had overlooked the point that reliability is not a property of a single score but of a group of scores. Livingston also indicated that the larger criterion-referenced reliability does imply a more dependable overall determination, when this decision is to be made for all individual scores in the distribution.

Meredith and Sabers (1972) also took issue with Livingston's concept of CRT reliability estimation as variability around the criterion score, pointing out that CRM is concerned primarily with the accuracy of the pass-fail decision and is relatively unconcerned with various levels of attainment above or below the criterion level.

Roudabush and Green (1972) presented several methods for arriving at reliability estimates for CRTs. The first involves ordering items hierarchically according to increasing difficulty. Roudabush and Green proposed that error of measurement is demonstrated if a student fails an easier item while passing a series of more difficult items. Oakland (1972) pointed out that it is very difficult to establish the needed hierarchical order. This objection has been raised since Guttman first (1944) proposed the technique of hierarchical ordering. Roudabush and Green's second technique used point-biserial correlations between parallel tests. Their results with this method

were far from encouraging (reported correlations were fairly low) and, in addition, there is difficulty inherent in the development of parallel tests. Their third method involved the use of regression equations to predict item criterion scores, but has not yet been fully explored.

Hambleton and Novick (1971) proposed regarding CRT reliability as the consistency of decision-making across parallel forms of the CRT or across repeated measures. They view validity as the accuracy of decision-making. This view departs from the classic psychometric view of reliability and validity. Hambelton and Novick view a decision-theoretic metric such as a "loss function" as being especially appropriate for use on CRTs. This metric serves to describe if an individual's true score is above or below a cutting score. The concept differs markedly from Livingston's (1972a) notion of regarding the criterion as the true score.

Swezey and Pearlstein (1974) suggested a simple technique for establishing the test-retest reliability of CRTs. The same group (of at least 30 people) is tested twice, close together in time. A four-fold table--first test administration, pass, fail--is then created, and a ϕ coefficient computed. It is important that the test group is not aware that it will be retested, so that practice does not occur between administrations. Swezey and Pearlstein proposed that a ϕ value of less than +.50 be considered indicative of questionable test reliability.

The importance of correct decision-making in CRT applications is also recognized by Edmonston, Randall, and Oakland (1972), who presented a CRT reliability model aimed at supporting decisions made during formative evaluation and at maximizing the probability of learning an established set of objectives. Criterion-referenced items are often binarily coded pass-

fail; therefore, summaries of group performance on two items of pre- and post-test can be displayed in a 2 x 2 contingency table. Edmonston et al. recommended utilizing cell proportions to provide information about the relationships between variables represented in the table. They found that a simple summation of the diagonal proportions $\sum_a paa$, provides a very useful measure of agreement between categories--where a is a method of indicating cells in a matrix and all cells have the same classification (pass-fail).

Thus, $\sum paa$ is used as a measure of association of cells in 2 x 2 tables, as opposed to chi-square used as a measure of independence. Thus, the coefficient of agreement is computed from cell values, rather than from marginal values, and should be used, according to Goodman and Kruskal (1954), for cases "in which the classes are the same for two polytomies. . . but differ in that assignment to class depends on which of two methods of assignment is used". Two test items, both of which are scored on a pass-fail basis, satisfy Goodman and Kruskal's conditions for the use of the coefficient of agreement.

They also recommended a supplemental measure λ_r (Lambda sub r) a variance-free coefficient. Goodman and Kruskal (1954) define λ_r:

$$\lambda_r = \frac{\sum paa - 1/2\ (PM\cdot + P\cdot M)}{1 - 1/2\ (PM\cdot + P\cdot M)}$$

where: PM· and P·M are the modal class frequencies for each of the two cross-classifications. λ_r may be interpreted as the relative reduction in the probability of error of classification when going from a no-information situation to an other-method-known situation. The no-information situation

refers to the probability of correctly classifying an individual randomly selected from a population on a dichotomous variable (e.g., able to perform item X, or not able to perform item X) when no information is known. The other-method-known situation refers to the probability of correct classification on item X when the classification from item Y is known, or vice-versa.

λ_r can take values from -1 to +1. A value of +1 indicates that both measures yield the same classification in all cases. A value of -1 indicates that the two measures never agree (no one who passes item 1 passes item 2, no one who fails item 1 fails item 2) and that the two modal frequencies of classification sum to unity. A difficulty with using λ_r is that if two measures are independent, λ_r has no set value--i.e., its value is not 0.

Edmonston et al. feel the reliability estimate most useful to CRM is the extent of temporal fluctuation. They suggested that, minimally, CRT items should provide stable estimates of knowledge of curriculum content; $\sum_a paa$ and r can be used to provide estimates of this stability. They recommend that $\sum_a paa$ be used to judge the re-test reliability of each item. However, when item re-test reliability falls below an arbitrary criterion (Edmonston et al. recommend 89%) and into a zone of decision, λ_r is employed as a descriptive measure of the amount of information gained by employing a second item (the re-test) in making curriculum or placement decisions. The method for making such decisions is not clear. Edmonston et al. stated only that "if knowledge of the retest score provides additional information as to how students can be classified, the item is retained."

In the same vein as Edmonston et al., Roudabush (1973) described reliability as the appropriateness of decisions affecting the treatment of examinees. Roudabush emphasized "minimizing risk or cost to examinee." The decision

is whether to discontinue instruction, remediate, or wash-out.

Validity

As is the case with NRT development, determination of validity for CRTs has seen less investigation than reliability. However, it is generally agreed that content validity is a paramount concern in CRT development. According to Popham and Husek (1969), content validity is determined by "a carefully made judgment, based on the test's apparent relevance to the behaviors legitimately inferable from those delimited by the criterion."

Klein and Kosecoff (1973) offered a distinction among the three most common methods of establishing content validity. In "systematic test development" the rationale of the systematic test development procedure used is explained, indicating why it should yield a content valid test. In the method of "expert judgment", content experts are given objectives and items, and are then asked to match items to objectives. The more accurately they can do this, the higher the content validity of the CRT using these items to measure the objectives. The third method uses "item analysis" to assess internal consistency "and/or see whether an item on a given objective correlates more highly with other items for this objective than it does with items on other objectives."

Klein and Kosecoff pointed out, however, that all three methods are limited by dangers involved with internal consistency techniques applied to CRM, and by possible lack of score variance. They noted, though, that lack of variance is a problem that "usually appears to be more theoretical than actual", and that "if enough students are tested, then one will discover sufficient variance in the levels of performance and/or in the time it takes

to achieve a given level".

McFann (1973) viewed the content validation of training as having two major dimensions. The first is the role of the human within a general operating system. Generally, this is defined by means of task analysis. The second dimension involves skills and knowledges a trainee brings with him to the course; training content can then be viewed as a residual of what must still be imparted to the trainee. The decision of what to include in training must also be tempered by management orientation to cost and effectiveness.

McFann stated that "decisions made on the units or procedures by which output (course completions) are to be evaluated, has an influence on validation of training content". McFann then elaborated, indicating that the decisions to which he referred include answers to questions such as:

> Will a normative or a fixed criterion approach be employed? What is the form of the evaluation? Will specific task performance be measured? Will transfer to other areas be emphasized? Will they be presented in a problem approach?

In other words, the way(s) in which one chooses to assess instructional outcomes will affect the validation of instructional procedures. In fact, McFann stated, "Answers to such questions will influence training content". McFann saw the validation of training content as a dynamic, interactive process, whereby training content is initially determined and then, on the basis of feedback about student performance on the job, instructional content as well as instructional methods are modified to improve overall effectiveness.

Edmonston, Randall, and Oakland (1972) held that content validation is

central to CRT development. CRT items are sampled from a theoretically large item domain, and must be representations of specified behavioral objectives.

The American Psychological Association's Standards for Educational & Psychological Tests (1974) discussed content validity, and noted that, "An employer cannot justify an employment test on grounds of content validity if he cannot demonstrate that the content universe includes all, or nearly all, important parts of the job". This APA document also discussed construct validity (which it found most applicable in research studies), and criterion-related validities. These latter include both concurrent and predictive validities. Criterion-related validities allow inference from test scores to standing on other, specified criteria. According to the APA standards, "Predictive validity involves a time interval during which something may happen. . . [while] concurrent validity reflects only the status quo at a particular time."

Swezey and Pearlstein (1974), have stated that content validity, "is probably the single best way of assessing whether or not your CRT measures what it is supposed to measure. . .[since it] is a matter of the extent to which a test corresponds with its objectives". But, they also noted that content validity can only be said to exist when a test consists of high-fidelity items, and that "Whether or not your test has content validity, you should also compute statistical estimates of concurrent validity, predictive validity, or both". Swezey and Pearlstein furnished simple techniques for computing concurrent and predictive validities, both of which employ the ϕ coefficient for correlating CRT results with appropriate, other measures

of the performance in question. A four-fold matrix (CRT, other measure, pass, fail) is analyzed via ϕ , and values less than +.50 are regarded as indicative of unacceptable concurrent or predictive validity. The technique only varies in that it is applied concurrently in the one case, and predictively (i.e., several months intervene between CRT and other measure) in the other case. Swezey and Pearlstein cautioned that the other measure must be suitable, and that the validation sample be representative and relatively large.

Hambleton and Novick (1971) proposed a validity theory in which a new test serves as criterion. Hambleton and Novick apply a decision-theoretic approach to both reliability and validity. Their suggested measure of reliability is "the proportion of times that the same decision would be made with the two parallel instruments". Hambleton and Novick indicated that a decision-theoretic approach to validity takes the same form "except--that a new test (Y) would serve as criterion and the qualifying score on the second test need not correspond with the qualifying score on the predictor CRT. The criterion 'test' might well be derived from performance on the next unit of instruction, or it could be a job-related performance criterion". Lack of correspondence between qualifying scores (i.e., cut-off points) does not necessarily make a predictor test invalid. This would be the case for norm-referenced measurement (NRM), but for CRM what is predicted is whether one will be above (or below) the qualifying score on a criterion test.

Although this approach appears reasonable, it seems that different conclusions may be reached if test Y were a job-related criterion as opposed to performance on the next unit of instruction. The different conclusions

could, however, yield approximations of convergent and divergent validity. Validation of a test determined by correlating it with another test may, however give a distinct overestimate of "validity". This is particularly true when the tasks on the two tests are similar.

Edmonston et al. (1972) advocated a method of CRT validation--which they termed a criterion-oriented approach--that includes both concurrent and predictive validity. In order to obtain complete information about an item and the objective it assesses, the relationship of a CRT to other measures should be considered (i.e., ratings by teachers or performance on suitable NRT measures). Edmonston et al. viewed these as measures of concurrent validity. In addressing problems of predictive validation, Edmonston et al. concurred with Kennedy (1972), proposing that tests of curriculum mastery (which represent higher order concepts taught within several curriculum units) be used as criteria against which unit test items would be assessed as to their predictive power. In addition, unit test items which are more temporally proximate should agree more strongly with Mastery Test items than items sequenced earlier. Final verification of this scheme of validity determination requires factorially pure items, and this may be a bit too much to ask of item writers.

Edmonston et al. endorsed an approach to construct validity initially put forth by Nunnally (1967). Nunnally pointed out that constructs are abstract variables, and that the more measures one obtains relating to a construct, the more explicitly defined that construct becomes. The "internal network" is, in Nunnally's terms, an <u>internal structure</u> based on the "correlations among the measures of observables in a particular set." This

internal structure may show that measures are related to the same thing--which is evidence that the set may be interpreted as a unitary construct. Or, it may turn out that the structure indicates "that two or more things are being measured by members of the set", in which case, a unitary construct is no longer sufficient. The example that Nunnally offered is that the internal structure of a set of measures purportedly measuring anxiety, is such that it is clear that two types of anxiety are actually being measured. Thus, it is appropriate to break the original set of anxiety measures into two sets, anxiety type one, and anxiety type two, corresponding to those variables which actually intercorrelate highly.

Nunnally concluded that "If all the correlations among members of the set are very low, it is illogical to continue speaking of the variables as constituting a set. . ." In Nunnally's view, the measurement and validation of a construct involve the determination of an internal network among a set of measures, and the consequent formation of a network of probability statements. This notion is similar to Cronbach and Meehl's (1955) enunciation of the need for a "nomological network" with which to validate a construct. Edmonston et al. indicated that the "specification of a hierarchy of learning sets among items would seem to be the ultimate goal of construct validation procedures, enabling the development of internal and cross structures between items and the consequent understanding of the inter-relationships of all curriculum areas". This concept would be difficult to implement, as the construction of learning sets is not an easy procedure. Also, difficulty can be expected in attempting to establish a network of relationships sufficient to completely define a construct.

In Roudabush's (1973) view of validity, CRT items are designed to sample the specified domain of behavior as purely as possible, and are then tried out to determine their sensitivity to instruction. A 2 x 2 contingency table containing post-test and pre-test outcomes is the basis for analysis:

Post-test

		-	+	
Pre-test	-	f_1	f_2	$f_1 + f_2$
	+	f_3	f_4	$f_3 + f_4$
		$f_1 + f_3$	$f_2 + f_4$	

f_1 = failed both pre- and post-

f_2 = failed pre-, passed post-

f_3 = passed pre-, failed post-

f_4 = passed both pre- and post-

Marks and Noll (1967) assumed that f_3 is due to guessing, and derived a sensitivity index named (s), which is simply the proportion of cases missing the item on the pre-test and passing it on the post-test, with a correction for guessing.

$$s = \frac{\hat{f}_2}{\hat{f}_1 + \hat{f}_2} \quad \text{where} \quad \hat{f}_2 = \frac{(f_2 - f_3) \quad (f_2 + f_1)}{f_1}$$

$$\hat{f}_1 = \frac{(f_3 + f_1)^2}{f_1}$$

Roudabush (1973), however, found that to derive a "reasonably reliable" value for the index, there should be at least 50 cases who missed the item at pre-test (f_1), while if the f_4 cell is high, the index will have little value (neither will the item). A problem here may be ensuring that different but parallel items are used for pre- and post-tests. This problem is a practical one, but is particularly acute when complex content domains are involved.

Guion (1974) explored the relationships between job-relatedness and several validities, and used employment tests--tests used to predict who is suitable for hiring and subsequent training--as an example. He suggested, "that an employment test may provide a basis for inferences that have criterion-related validity, or construct validity, or content validity, or all of these, and still not be job related". Guion viewed job-relatedness "as the extent to which the hypothesis of a relationship between the hiring requirement and job behavior can be accepted as logical". Guion concluded that one new technique that might help improve psychological measurement, bridging the job relatedness--validities chasm "is the content-referenced measurement of mastery".

These treatments of CRT validity all exhibit difficulties that might prove insurmountable to a test constructor dealing with "real world" problems. Content validity however, is of primary importance in CRM and can be reasonably ensured by careful attention to objective development. Construct validity will probably prove elusive, if only due to the complexity of operations and measures required for its demonstration. Predictive and concurrent validities appear practicable in many situations.

Construction Methodology

NRTs are designed primarily to measure individual differences. The meaning which can be attached to a score depends upon a comparison of that score to a relevant norm distribution. A NRT is constructed to maximize test score variability, since such a test is likely to produce less errors in ordering individuals on the measured ability. Since NRTs are often used for selection and classification purposes, minimizing errors of ordering is extremely important.

NRTs are constructed using traditional item analysis procedures. It is partly because of this that test scores cannot be interpreted relative to some well-defined content domain. That is, items are selected to produce tests with desired statistical properties (e.g., difficulty levels around .5), rather than to be representative of a content domain.

CRTs, on the other hand, tend to have restricted ranges of variance. Thus, they are not easily subjected to traditional item analysis procedures. There are, however, ways of obtaining the necessary range of variance. Haladyna (1974) performed a study to demonstrate the feasibility of combining pre- and post-instruction CRT scores in order to increase score variance, thereby permitting the use of classical psychometric methodology for item analysis and test reliability. He administered units of instruction and CRTs to 189 undergraduate education students, and computed test and item statistics for three samples:

1. Preinstruction students, representing a nonmastery population,
2. Post instruction students, representing a mastery population, and
3. The above two samples combined.

Haladyna found that combining the samples greatly increased variance over either of the samples alone, and that point biserial discrimination indexes computed for the combined samples appeared to be "the most efficient method for obtaining information about the adequacy of the CR test items".

Woodson (1974a and 1974b) has argued that item (and test) variance is a necessary condition for item selection in CRT development, as well as in NRT development. He noted that variance in NRM results from observations on random samples of individuals in a population, while CRM variance results from observations on a sample representative of the range of the characteristic measured. This range of the characteristic can vary from no one passing any items (as in pre-instruction testing) to everyone passing all items (as in the ideal postinstructional outcome). Noting that "The better an item discriminates among instances of the characteristic within the range of interest, the more information the item gives us", Woodson concluded that "If our measurement devices are sufficiently precise, individuals will be ordered on an appropriate scale". Woodson's concept of CRT variance, and its value in developing CRTs capable of ordering individuals, has yet to be empirically verified, though.

Item homogeneity is also much sought in development of NRTs. The ultimate purpose is to spread out individuals by maximizing the discriminating power of each item. The emphasis is on comparing an individual's response to the responses of others. The interest is not in absolute measurement of individual skills, as in CRTs, but only in relative comparison. Thus, item homogeneity is not directly applicable to CRM.

Nevertheless, item analysis is an important tool in test construction and therefore has application to the construction of CRTs. Although content validity is an important characteristic for a CRT item, other considerations having to do with sensitivity and discriminating power of an item are also important. These features are important in evaluating instruction and in ensuring correct decisions about an individual's progress through instruction.

In CRT development, an item difficulty index may be useful for selecting items. However, item difficulty is used differently than in NRM. If the content domain is carefully specified, test items written to measure accomplishment of an objective should also be carefully specified and closely associated with the objective. Thus, all items associated with the same objective should be answered correctly by approximately the same proportion of examinees in a group. Items which differ greatly should be examined carefully to determine if they coincide with the intent of the objectives.

Similarly, item discrimination indexes can be useful in CRT development. Negative discrimination indexes warn that CRT items need modification, or that the instructional process is faulty. A negative index would be indicative of a high proportion of "false negatives".

Klein and Kosecoff (1973) discussed item analysis as a means for improving CRT item quality, and noted that selection of "good" items varies as a function of item analysis method. They then described two concepts underlying four general types of item analysis methods used in the development of CRTs:

1. Sensitivity to instruction--"good" items are failed prior to the relevant instruction, and passed following instruction, and
2. Discrimination among items on the basis of internal consistency--"good" items discriminate between those who do well on the test as a whole (or on some external criterion) and those who don't.

Klein and Kosecoff delineated the four general item analytic approaches based on these two concepts: Comparison group (masters versus non-masters, or have-received-instruction versus have-not-received-instruction); single group using pre- and post-instruction tests; single group using posttest only; and single group with repeated measures (test is administered until mastery is achieved on all items, and pass-fail patterns are examined to detect reversals). Klein and Kosecoff's analysis indicated that the latter two methods are less applicable than the first two. They also cautioned that, when the first type of method is used, both groups must be equated as to general intellectual ability, or as to other factors that might contaminate the comparison.

One attempt to use item analysis techniques to develop test evaluation indexes was undertaken by Ivens (1970). Ivens has defined reliability indexes based on the concept of within-subject score equivalence. Item reliability is defined as the proportion of subjects whose item scores are the same on the post-test, as on either a retest or a parallel form. Score reliability is then defined as the average item reliability.

Rahmlow, Matthews, and Jung (1970) suggested that the function of a discrimination index in a CRT is primarily that of indicating item homogeneity with respect to the specific instructional objective measured. These authors

focused on shifts in item difficulty from pre-instruction to post-instruction

Helmstadter (1972) compared the following alternative indexes of item usefulness:

1. Item discrimination based on high and low groups on a post-instructional measure.
2. Shift in item difficulty from pre- to post-instruction.
3. Item discrimination based on pre- and post-test performance.

Shift in item difficulty from pre- to post-instruction produced results significantly more similar to the pre-post discrimination index, than did the high-low group, post-test discrimination index comparison.

Helmstadter also compared an item discrimination index applied to pre- and post-instruction with difficulty indexes derived in the same fashion. His findings resulted in the conclusion that caution should be observed when using traditional item analysis procedures with CRTs. In a similar finding, Roudabush (1973) described a situation where use of traditional item statistics would have resulted in some objectives being over-represented while others would not be represented at all.

Ozenne (1971) has developed an elaborate model of subject response which he used to derive an index of sensitivity. In this formulation, the sensitivity of a group of comparable measures, given to a sample of subjects before and after instruction, is defined as the variance due to the instructional effect divided by the sum of the variance due to the instructional effect and error variance. This index was however, developed for a severely restricted sample in order to allow an analysis of variance treatment. Further development is indicated before the technique has general usefulness for sensitivity measurement or item selection.

New procedures have been developed for item analysis of specific CRTs, but evidence as to their generalizability is lacking. If item analytic procedures are to be used in evaluating CRTs, one must consider what sort of score is produced by the item. The most typical scoring involves a pass-fail dichotomy. A CRT item can result in two types of incorrect decisions. Roudabush and Green (1972) referred to these errors as "false positives" and "false negatives". In this view, reliability is concerned with the CRT's ability to consistently make the same decision. Consequently, validity becomes the ability of a CRT to make the "right" decision (i.e., avoiding false negatives and false positives). In these authors' view, the adequacy of a CRT is determined by its ability to discriminate consistently and appropriately over large numbers of items.

Swezey and Pearlstein (1974) suggested comparing "masters" and "non-masters" as to pass-fail on items, thereby circumventing the internal consistency problem. "Masters" and "non-masters" can be defined either in terms of completion/noncompletion of the relevant instruction, or in terms of skill level on some external criterion; i.e., the "master" has had considerable experience in the subject area, while the non-master has not. A ϕ coefficient is computed for each item ("master-nonmaster", pass-fail), and a value of less than +.30 indicates an item of questionable utility.

Carver (1970) proposed two procedures to assess reliability of CRT items. For a single form, he suggested comparing the percentage meeting criterion level in one group to the same percentage in another "similar" group. For homogeneous sets he recommended using one group and comparing the percentages who meet the criterion on all items.

Meredith and Sabers (1972) pointed out that the way in which two CRT items, whether identical or parallel, identify the same individual with regard to his attainment of criterion level must be determined. With regard to item analysis procedures, if a CRT item is administered before and after instruction, and does not discriminate, there are alternatives to labeling it unreliable. A non-discriminating item may simply be an invalid measure of the objective or it may indicate that the instruction itself is inadequate or unnecessary. Meredith and Sabers suggested the use of a matrix consisting of the pass-fail decisions of two CRTs. By defining two CRT items as being the same measures, it is possible to examine test/re-test reliability. Without time intervening between the measures however, reliability is of the concurrent variety. In addition, problems exist with the acceptability of defining two CRTs as the same. Considerable confusion is evidenced in the use of "same" and parallel forms without formal definitions. Similarly, it was stated that if one CRT item is a "criterion measure", then the validity of the other CRT can be determined. By definition, both are criterion measures, and if one is external to the instructional domain, then it is not a CRT item in the same sense. Various coefficients were presented, but difficulties in definition limit their usefulness.

Fidelity

Fredericksen (1962) has proposed a hierarchical model for describing levels of fidelity in performance evaluation. The model uses six categories:

1. Solicit opinions. This category, the lowest level, may often miss a crucial question (e.g., to what extent has the behavior of trainees been modified as a function of the instructional process?).

2. Administer attitude scales. This technique, although psychometrically refined via the work of Thurstone, Likert, Guttman, and others, assesses primarily a psychological concept (attitude) which can only be presumed to be concomitant with performance.
3. Measure knowledge. This is the most commonly used method of assessing achievement. This technique is usually considered adequate only if the training objective is to teach knowledge, or if highly defined, fixed procedure tasks are involved.
4. Elicit related behavior. This approach is often used in situations where practicality dictates observation of behavior thought to be logically related to the criterion behavior.
5. Elicit "What Would I Do" behavior. This method involves presentation of brief descriptions or scenarios of problem situations under simulated predesigned conditions; the subject is required to indicate how he would solve the problem if he were in the situation.
6. Elicit lifelike behavior. Assessment under conditions which approach the realism of the real situation.

Measurement at any of these six levels possesses both advantages and disadvantages. An optimal solution would be to assess individual performance at the highest possible level of fidelity. Unfortunately, deriving observed performance data may involve a subjective (rating) scale, thereby requiring a subjectivity vs. fidelity tradeoff. In order to minimize subjectivity, it may be necessary to decrease the level of fidelity so that more objective measurements (such as time and errors) can be obtained. These measures can be conceptualized as surrogates that in some sense embody real criteria,

but have the virtue of measurability (Rapp et al., 1970). An actual increase in overall criterion adequacy may result from a gain in objectivity which compensates for a corresponding loss in fidelity.

The question of fidelity addresses the issue of how much the test should resemble actual performance. Fidelity is not usually at issue in NRM and has its primary application in criterion-referenced performance tests.

There are often trades to be made between fidelity and cost. But a more salient issue is how to modify fidelity to satisfy needs of the testing situation, while retaining the essential stimuli and demand characteristics of the actual performance situation. Swezey and Pearlstein (1974) suggested that when creating items, the CRT developer "Select the format that best approximates the behavior specified by the objective", and that "will permit the highest level of fidelity practicable".

Osborn (1970) addressed problems of finding efficient alternatives to work sample tests. Osborn was concerned with developing a methodology to allow derivation of cheaper procedures while preserving content validity. There are many situations where job sample tests are not feasible, and job-knowledge tests are not relevant. The existance of intermediate or "synthetic" measures would be a great boon to evaluating performance in these situations; however, specific methods for developing such measures are lacking.

Osborn gave a brief outline of a method for developing synthetic measures. He presented a two-way matrix defined by methods of testing terminal performance (simple to complex) and component (enabling) behaviors. This matrix serves as a decision-making aid by allowing the test constructor to choose the most cost-effective test method for each behavior. Tradeoffs must be made by

the test constructor among test relevance, obtaining diagnostic performance data, ease of administration, and cost. Osborn's notions are intriguing but more developmental work is needed before a workable method for deriving synthetic performance tests is available.

Vineberg and Taylor (1972) addressed a topic close to the fidelity issue: the extent to which job knowledge tests can be substituted for performance tests. Practical considerations have often dictated the use of paper-and-pencil job knowledge tests because they are simple and economical to administer and easy to score. However, the use of paper-and-pencil tests to provide indexes of performance is considered to be poor practice.

HumRRO research under Work Unit UTILITY compared the proficiency of Army men at different ability levels and with different amounts of job experience. This work provided Vineberg and Taylor the opportunity to examine relationships among job sample test scores and job knowledge test scores in four U.S. Army jobs that varied greatly in type and complexity. Vineberg and Taylor found that job knowledge tests are valid for measuring proficiency in jobs where" (1) skill components are minimal, and (2) job knowledge tests are carefully constructed to measure only information directly relevant to performing the job at hand. Given the high costs of obtaining performance data, these findings indicate that job knowledge tests are indicated where careful job analysis has determined that skill requirements are minimal.

In a similar study, Engel and Rehder (1970) compared peer ratings, a job knowledge test, and a work-sample test. While the knowledge test was acceptably reliable, it lacked validity, and reading ability tended to enter into the score. Peer ratings were judged to have unacceptable validity and

were essentially uncorrelated with the written test. The troubleshooting items on the written test exhibited a moderate level of validity, while the corrective-action items had little validity. Finally, Engel and Rehder noted that the work-sample test is the most costly and difficult to administer, while peer ratings and written tests were less costly and easier to administer.

Process vs. Product Measurement

Osborn (1973a and 1973b) discussed an important topic directly related to CRT validity and fidelity. Osborn pointed out that task outcomes and products are often used to assess student performance while measures of how the tasks are done (processes) generally pertain to the diagnosis of instructional systems. Time or cost factors sometimes preclude the use of product measures, thus leaving process measures as the only available criteria. There are cases where this focus on process is legitimate and useful, but many where it is not. Osborn developed three classes of tasks to illustrate what the relative roles of product and process measurement should be:

"1. Tasks where the product is the process.

2. Tasks in which the product always follows from the process.

3. Tasks in which the product may follow from the process."

A relatively few tasks can be classified as the first type. Osborn offered gymnastic exercises and springboard diving as examples. More tasks belong to the second classification, fixed procedure tasks. In these tasks, if the process is performed correctly, the product follows. The largest single class of tasks is of the third type. For tasks of this last type, the process may appear to have been correctly carried out for cases in which the product was not attained. Osborn offered two reasons for this: either,

(1) "we are unable to fully specify the necessary and sufficient steps in task performance", or (2) "we do not or cannot accurately measure them". An example of aim-firing a rifle was given as an illustration that there is no guarantee of acceptable markmanship even if all procedures are followed. In this case, process measurement is not an adequate substitute for product measurement.

For tasks of the first two types, Osborn concludes that it really doesn't matter which measure is used to assess proficiency. But for tasks of the third type, product measurement is indicated. There are however, a number of type 3 tasks where product measurement is impractical because of cost, danger, or other constraints. In these cases, process measures are substituted, with a resulting decrement to the validity of the measure. Osborn poses a salient question for the test developer: "If I use only a process measure to test a man's achievement on a task, how certain can I be from this process score that he would also be able to achieve the product or outcome of the task?" Osborn holds that "Where the degree of certainty is substantially less than that expected by errors of measurement, the test developer should pause and reconsider ways in which time and resource limitations could be compromised in achieving an approximation to product measurement". Osborn concluded by noting: "The accomplishment of product measurement is not always a simple matter; but it is a demanding and essential goal to be pursued by the performance test developer if his products are to be relevant to real world behavior."

Swezey and Pearlstein (1974) have also addressed process versus product measurement, and assist versus non-interference methods of scoring in CRT

development. They recommended process measurement in addition to, or instead of, product measurement when: Diagnostic information is desired; critical points in the process, if misperformed, may cause injury or damage; additional scores are needed on a particular task; the product always follows from the process; and there is no product at the end of the process.

Another issue important to the construction of complex CRTs, is bandwidth fidelity (Cronbach and Gleser, 1965), i.e., the question of whether to obtain precise information about a small number of competencies or less precise information about a larger number. Hambelton and Novick (1971) conclude that the problem of fixing the lengths of subscales to maximize percentage of correct decisions on the basis of test results, has yet to be resolved or even satisfactorily defined.

Issues Related to CRT Construction

Although construction methodology for NRTs is well-established and highly specified, construction techniques for CRTs have been less well-specified. There have been, however, several attempts to formalize the CRT construction process. Ebel (1962) describes the development of a criterion-referenced test concerning knowledge of word meanings. Three steps were involved:

1. Specification of the universe to which generalization is desired.
2. A systematic plan for sampling from the universe.
3. A standardized method of item development.

These characteristics together serve to define the meaning of test scores.

Flanagan (1962) indicates that a variant of Ebel's procedure was used in project TALENT. Tests used in the areas of spelling, vocabulary, and

reading were not based on specific objectives, but developed by systematically sampling a relevant domain. Fremer and Anastasio (1969) also put forth a method for systematically generating spelling items from a specified domain.

Osburn (1968) notes two prerequisites for inferences drawn about a domain of knowledge from performance on a collection of items:

1. All items that could possibly appear on a test should be specified in advance.
2. The items in a particular test should be selected at random from the content universe.

It is rarely feasible to satisfy the first prerequisite for complex behavior domains. However, the problem of testing all items can be overcome, at least in highly-specified content areas, by the use of an <u>item form</u> (Hively, 1968, 1973; Osburn, 1968). The item form generally has the following characteristics (Osburn, 1968):

1. It generates items with a fixed syntactical structure.
2. It contains one or more variable elements.
3. It defines a class of item sentences by specifying the replacement sets for the variable elements.

But Klein and Kosecoff (1973) have noted that even very specific objectives--e.g., "compute the correct product of two single digit numerals greater than 0, the product not exceeding 20"--may yield possible item pools "of well over several thousand items". In the example objective above, there are 29 pairs of possible numerals times at least 10 different suitable item types (e.g. 6 x 2, vs $\underline{\overset{6}{x}2}$, vs (6)(2), vs 6 x _ = 12, etc.) times variations in numeral sequence (e.g., 6 x 2, 2 x 6) times variations in item format (e.g., multiple

choice, fill-in-the-blank, etc.) times variations in presentation mode (e.g., oral or written) times variations in mode of response (oral or written), equaling 4,640 potential items, assuming only two types for each of the variations indicated. So, item forms require careful consideration, even for highly defined areas such as mathematics.

Shoemaker and Osburn (1969) describe a computer program capable of generating both random and stratified random parallel tests from a well-defined and rule-bound population. Results have led to the conclusion that difficulties in defining test construction processes are directly related to the complexity of the behavior the test is designed to assess (Jackson, 1970). Where the domain is easily specified (as in spelling) the construction process is simplified. Jackson (1970) concludes, "For complex behavior domains, it appears that at least until explicit models stated in measurable terms are developed, a degree of subjectivity in test construction (and attendant population-referenced scaling) will be required." The best approach appears to be the use of a detailed test specification which relates test item development processes to behavior.

Edgerton (1974) has suggested that the relationships among instructional methods, course content, and item format have not been adequately explored. Item format should require thinking and/or performing in the patterns sought by the instructional methods. If the instruction is aimed at problem solving, then the items should address problem solving tasks and not, for example, knowledge about the required background content. Edgerton suggests that if one mixes styles of items in the same test, one runs the risk of measuring "test taking skill" instead of subject matter competence. In a practical

application, Osborn (1973) suggests a fourteen-step procedure in the course of developing tests for training evaluation. Swezey and Pearlstein (1974) have suggested that the following factors must be considered when designing a test plan to guide item development:

1. Overcoming practical constraints in test administration (time, manpower, and facilities availability, etc.) by selecting among objectives (randomly testing non-critical objectives) or modifying objectives.
2. Planning item format and level of fidelity.
3. Sampling items within objectives.
4. Sampling among multiple conditions.
5. Deciding how many items to include on the test.

Mastery Learning

Besel (1973 a, b) contends that norm-group performance data are useful for the construction of CRTs. Besel defines a CRT as a set of items sampled from a domain which is judged to be an adequate representation of an instructional objective. The domain should be described in sufficient detail to allow independent test developers to generate equivalent items measuring the same content in an equally reliable fashion.

Besel recommends a "Mastery Learning Test Model" to provide an appropriate algorithm for support of mastery/non-mastery decisions. The Model and its underlying true score theory, is related to a notion developed by Emrick (1971). Emrick assumed that measurement error was attributable to two sources: α, the probability that a non-master will correctly answer an item ("false positive") and β, the probability that a master will give an incorrect answer

("false negative"). Emrick's model assumes that all item difficulties and inter-item correlations are equal, a somewhat difficult assumption. Besel (1973 a, b) developed algorithms for estimating α and B. Three data sources are required:

1. Item difficulties
2. Inter-item covariance
3. Score histograms

Besel reports that "the usage of an independent estimate of the proportion of students reaching mastery resulted in improved stability of Mastery Learning parameters" in a tryout sample. Improved stability of α and B should promote increased confidence in mastery/non-mastery decisions. Besel's computational procedures are however, quite involved, using a multiple regression approach which requires independent a priori estimates of variance due to conditions. Besel also points out that B is estimated best for a group when the mastery level is lowered, while the reverse is true for α. In other words, Besel has empirically established a relationship between errors of misclassification and criterion level. A decision, however, has not been made concerning the relative cost/effectiveness of competing errors of misclassification. Such decisions may be specific to instructional situations.

Establishing and Classifying Instructional Objectives

The development of student performance objectives for instructional programs has become a widespread process within the educational community. Information is generally derived from instructional objectives, which provide not only specifications for instruction, but also the basis for instructional evaluation (Lyons, 1972). Ammerman and Melching (1966) trace the interest in behaviorally-stated objectives from three independent movements within

education. The first derives from the work of Tyler (1934, 1950, 1964) and his associates, who worked for over 35 years at specifying goals of education in terms of meaningful and useful information for the classroom teacher. Tyler's work has had considerable impact in the trend toward describing objectives in terms of instructional outcomes.

The second development came from the need to specify man-machine interactions in modern defense equipment configurations. Miller (1962) was responsible for pioneering efforts in describing and analyzing job tasks. Chenzoff (1964) reviewed the then existing methods in detail, and many more have appeared since that date. More recently Davies (1973) classified task analysis schemes into six categories:

1. Task analysis based upon objectives, which involves analysis of a task in terms of the behaviors required.
2. Task analysis based upon behavioral analysis of concepts, chains, etc.
3. Task analysis based upon information processing needs for performance.
4. Task analysis based upon a decision paradigm which emphasizes the judgment and decision-making rationale of the task.
5. Task analysis based upon the subject matter structure of a task.
6. Task analysis based upon vocational schmeatics which involve analysis of jobs, duties, tasks and task elements.

The point of Davies' breakdown is that there is no single task analysis procedure which is always applicable. The typical approach is to create a new task analysis scheme or modify an existing scheme to suit the needs of the situation at hand.

The third development was the concept of programmed instruction, which

required writers of programs to acquire specific information on instructional objectives.

It is apparent that the use of instructional objectives has now become an accepted educational practice. A critical event in this process was the publication of Mager's (1962) little book Preparing Instructional Objectives. In this work, Mager set forth requirements for the form of a useful objective, but did not deal with procedures by which one could obtain information to support preparation of the objectives. A series of additional works, including one on measuring instructional intent (Mager, 1972), have dealt more thoroughly with such issues.

Actual behaviors exhibited by acceptable performers are generally preferred as the bases for constructing instructional objectives. However, data can come from a variety of sources, including:

1. Supervisor interview
2. Job incumbent interview
3. Direct observation of performance
4. Inferences based upon system operation
5. Analysis of "real world" use of instruction
6. Instructor interview

Many sophisticated methods are used to derive these data. Flanagan's (1949) "critical incident technique", and the modifications it has inspired, are good examples of efforts aimed at identifying essential performances, while eliminating information not directly related to the successful accomplishment of a job-related task.

The choice of a method for deriving job training content must be based upon the type of performance, and upon other realistic factors such as

assessibility of the performance to direct observation. Generally the solution is less than ideal, but techniques such as Ammerman and Melching's (1966) can be used to review objectives and to provide a useful critique of the data collection method. An exhaustive review of the techniques for deriving instructional objectives is inappropriate here. The reader is directed to Lindvall (1964) and to Smith (1964) for comprehensive treatments of this question.

Klein and Koseoff (1973) have summarized four general procedures used in developing objectives for CRTs. The first, "expert judgment", is the most common approach in their opinion, and involves a small group of subject matter experts meeting to arrive at a consensus of important objectives to measure. Objectives thereby identified are screened on the basis of practical constraints, and are modified as necessary. The second procedure, "consensus judgment", is similar to the first, but uses "various groups such as community representatives, curriculum experts", etc. to decide which objectives should be measured. Appropriate measurement or curriculum personnel then translate the objectives into terms permitting assessment.

"Curriculum analysis", the third approach, involves analysis of curriculum materials (e.g., textbooks), and subsequent identification or inference of objectives therein by a team of curriculum experts. Finally, the fourth approach, "analysis of the area to be tested" is similar to the task analytic approaches previously discussed. Contents and behaviors in the subject area are identified, and organized hierarchically (or according to some other sequence) to derive objectives.

Ammerman and Melching (1966) have developed a system for the analysis and classification of terminal performance objectives. They examined a great number of objectives from diverse sources, and concluded that five factors accounted for the significant ways in which most existing performance objectives differed. These factors are:

1. Type of performance unit
2. Extent of action description
3. Relevancy of student action
4. Completeness of structural components
5. Precision of each structural component

Ammerman and Melching have identified a number of levels under each of these factors. For instance, factor #1 has three levels, from specific task, which involves one well-defined particular activity in a specific work situation, to generalized behavior, which refers to a general measure of performance, or way of behaving, such as the "work ethic".

With these five factors, and their sub-levels, it is possible to classify any terminal objective via a five digit number. This scheme is valuable for management control and for review of terminal performance objectives. Ammerman and Melching feel the method can fulfill three main purposes:

1. Provision of guidance for the derivation of objectives and for standardization of statements of objectives, so that all may meet the criteria of explicitness, relevance, and clarity.
2. Evaluating the proportion of objectives dealing with specific or generalized action situations.
3. Evaluating the worth of a particular method for deriving objectives.

This is a useful method, particularly where a panel of judges is available to review each objective. A coefficient of congruence can be computed among judges' placement of objectives on the five dimensions. Used in this fashion, the Ammerman and Melching method should prove useful in the development of instructional systems.

Developing Test Materials and Item Sampling

Hively and his associates (1968, 1973) have provided a useful scheme for writing items which are congruent with a criterion. Hively's efforts have been primarily in the area of domain-referenced achievement testing. In this system, an item form constitutes a complete set of rules for generating a domain of test items which are accurate measures of an objective.

Popham (1970) has pointed out that the item form approach has met with success in content areas having well-defined limits. In such areas (e.g., mathematics), independent judges tend to agree on whether or not a given item is congruent with the highly-specific behavior domain referenced by the item form. As less well-defined fields are considered, however, it becomes more difficult to prepare item forms so that they yield test items which are judged congruent with a given instructional objective. Popham (1970) has remarked: "Perhaps the best approach to developing adequate criterion-referenced test items will be to sharpen our skill in developing item forms which are parsimonious but also permit the production of high congruency test items."

Cronbach (1963, 1972) presents a generalizability theoretic approach to achievement testing. Cronbach's theory involves a mathematical model in the framework of which, an achievement test is assumed to be a sample from a large, well-defined domain of items. Parallel test forms are obtained

by repeated sampling. Analysis of variance techniques (particularly intra-class correlation) are used to obtain estimates of variance components due to sampling error, testing conditions, and other sources which may affect the reliability of scores.

Generalizability theory has been extended (Osburn, 1968) by including concepts of task analysis to allow sorting subject matter into behavioral classes. Osburn (1968) has termed this "Universe-defined achievement testing." Hively (1968, 1973) has used these techniques in an exploration of a mathematics curriculum. Mathematics represents a subject domain particularly suited to this approach and Hively reported success as evidenced by high intra-class correlations between sets of items sampled from a universe. The technique appears to have diagnostic utility, and is also relevant for examining relationships between knowledges and skills.

Rogers (1965) has stated that "The major problem in developing the test item is to clearly communicate the question or problem to the student". He suggested 11 practical guidelines to help surmount this problem, many of which entail logistical considerations in the presentation of performance test items to examinees.

Swezey and Pearlstein (1974) suggested that items be developed in conjunction with a carefully-defined test plan (See "Issues Related to CRT Construction" section). They also offered the following suggestions for the development of CRT items:

1. "Make the test items include the same conditions and standards (no more, no less) as those specified in the objective."
2. "Use graphs, drawings, and photographs when necessary for clear

communication."

3. "Present the test so it does not give the student hints as to the correct answer, but never make it extremely difficult simply to ensure a certain number of failures."

4. "Include necessary specific instructions with items".

They noted that items should be assessed for adequacy prior to submission for item analysis try-outs. Such assessments include making sure the items match objectives as to performance, conditions, and standards; the items are clear, unambiguous, and reasonably easy to administer; and that they are at the appropriate fidelity level, as determined previously.

Quality Assurance

According to Hanson and Berger (1971), quality assurance is viewed as a means for maintaining desired performance levels during the operation of a large-scale instructional program. Six major components in a Quality Assurance program are identified:

1. Specification of indicator variables. These are variables which measure important aspects of a program and are individually defined for each instructional system. Examples are:
   a. Pacing - measure of instructional time
   b. Performance - interim measures of learning, e.g., unit tests, module tests, etc.
   c. Logistics - indicator reports of failure to deliver materials, and other implementation difficulties resulting from poorly planned logistics.

2. Definition of decision rules. The emphasis here should be on indicators which signal major program failures. Critical levels may be determined on the basis of evidence from developmental work or from an analysis of program needs.
3. Sampling procedures. Questions about sampling procedures must be answered on the basis of an analysis of the severity of effects resulting from insufficient information. Factors to be considered include:
   a. Number of program participants to provide data
   b. How to allocate sampling units
   c. Amount of information from each participant
4. Collecting quality assurance data. Special problems concern the willingness of participants to cooperate in the data gathering effort. Data must be timely and complete. Hanson and Berger suggest a number of ways to reduce data collection problems:
   a. Minimize the burden on each participant by collecting only required data.
   b. Use thoroughly designed forms and simplified collection procedures.
   c. Include indicators which can be gathered routinely and without special effort.
5. Analysis and summarization of data. Some data may be analyzed as they come in; other data may have to be compiled for later analysis. The exact technique will depend on the type of decision the data must support.

6. Specification of actions to be taken. This step describes actions to be taken in the event of a major program failure. Alternatives should be generated and scaled according to severity of failure. Information on actions taken to correct program failures should be fed back into the program development cycle. Such feedback is an important source of guidance for program revision.

Hanson and Berger offer an illustrative example of how this process might be implemented. They note that quality assurance, as applied to criterion-referenced programs, acts to ensure that specified performance levels are maintained throughout the life of a program. If internal quality assurance programs of this sort are built into instruction, then the probability of an instructional program becoming "derailed" while functioning is minimized.

Designing for Evaluation and Diagnosis

Baker (1972) feels that the critical factor in instruction is not how test results are portrayed (NRT or CRT) but now they are obtained and what they represent. Baker suggests the term construct-referenced to describe achievement tests consisting of a wide variety of item types and well-sampled content ranges. These tests are generally of the norm-referenced type. Criterion-referenced tests, Baker feels, are probably better termed domain-referenced tests (see discussion of Hively et al., 1968, 1973). A domain specifies both the performance a learner is to demonstrate, and the content domain to which the performance is to generalize.

Baker uses the term objective-referenced test to refer to another subset of CRM. Objective-referenced tests start with objectives based upon observable behaviors from which it is possible to produce homogeneous items

relating to the objective. Baker feels the notion of domain-referenced tests is more useful than the notion of objective-referenced tests.

Each type of test provides different information to guide improvement on instructional systems. Construct-referenced tests provide information regarding the full range of contents and behaviors relevant to a particular construct. Objective-referenced tests provide items which exhibit similar response requirements relating to vaguely defined content areas. Domain-referenced tests include items which conform to a particular response segment, as well as to a class of content to which the performance is presumed to generalize.

According to Baker (1972), a test should ideally be capable of yielding information needed to implement an instructional improvement cycle. An ideal test should yield data on:

1. Applicable student abilities
2. Deficiencies in student achievement
3. Possible explanations for deficiencies
4. Alternative remedial sequences
5. Facility with which remedial sequences can be implemented.

All three types of tests provide useful data concerning student abilities. Construct-referenced tests are probably the most readily available, but are often not administerable on a cycle compatible with diagnosis, and are usually reported in a nomothetic manner. A well-designed objective-referenced test may be scheduled in a more useful fashion. Domain-referenced tests provide enabling information to allow instructors to identify areas in which students are competent. Identification of performance deficiencies is theoretically possible with all three sets of data. However, since cut-offs are often

arbitrary, none of the three tests will necessarily provide adequate information of this type. In addition, incentives are lacking since most accountability programs are used to punish deficiency rather than to promote efficiency.

Of the three test types, the domain-referenced tests give program developers the most assistance, for they provide clear information about appropriate types of practice items in the areas of content and performance measured by the test. However, Baker points out that domain-referenced items are hard to prepare, mainly because many content areas are not analyzed in a fashion to allow precise specification of the behaviors in the domain.

Establishing Cut-Off Scores

Prager et al. (1972) discuss cut-off points in mastery testing, and suggest that two general approaches exist. The first involves setting an arbitrary overall mastery level. A trainee either attains this criterion level or not. The second procedure requires that trainees attain the same mastery level in a given objective, but allows the levels to vary from objective to objective, depending upon difficulty of material, importance of the objective for later successful performance, etc. This second method seems more reflective of reality but as Prager et al. (1972) point out, it is also more difficult to implement, and to justify specific levels that have been decided upon. Prager et al. believe that for handicapped children at least, it is appropriate to set mastery levels for each child relative to his potential. Nitko (1971) concurs and suggests different cut-offs seems doubtful.

Lyons (1972) points out that standards must take into account the varying criticality of tasks. The criticality of a task is basically an assessment of the effect upon an operating system of incorrect performance of that task.

Criticality must be determined during task analysis, and must be incorporated into the training objective. Unfortunately, in many cases, task criticality is not an absolute judgment, and the selection of a metric for criticality becomes arbitrary.

The approach to reliability advocated by Livingston (1972) holds some promise for determining pass-fail scores. If Livingston's assumptions are accepted, then it becomes possible to obtain increased measurement reliability by varying the criterion score. If the criterion score is set so that a very high (or low) proportion passes, then one obtains reliable measurement. Unfortunately, it is not often possible to manipulate criterion scores to this extent. The training system may require a certain number passing, and the criterion score is frequently adjusted to provide the required number.

Graham (1974) compared existing eighth-grade, objective-based mathematics tests to domain-referenced tests designed to assess achievement on the same objectives, after both tests were administered to 151 eighth-grade students. He found that slight changes in item form introduced concommitant skills in addition to those specified by the objectives. These additional skills confounded the measurement of primary, objective-specified skills, and that "confounding increases the number of scores falling in the middle of the possible range. . ." This, in turn, increases the amount of overlap between the distributions of scores for masters and non-masters, thereby increasing both the number of scores "at or near any selected mastery cut-off score", and the likelihood of misclassification. So, for tests consisting of heterogeneous items--those in which measurement of several skills may be confounded--classification of masters and non-masters may be seriously affected by the

cut-off score selected: The score distribution for all examinees is likely to be rectangular. But, for tests composed of homogeneous items, the score distribution is more likely to be bimodal--masters and non-masters clearly separated--and there is much more latitude in the placement of the cut-off score before classification is affected.

Swezey and Pearlstein (1974) concurred with Graham's findings, stating "The more complex the skills assessed by the CRT and the more varied the type of performance or product, the greater is the danger of misclassification. . ." They also noted that immediate manpower needs and criticality of objectives must also influence placement of the cutting score, justifying lowering or raising the cut-off level, respectively. Finally, they stated emphatically that "If a test is measuring more than one objective, and cut-off scores are necessary, a cut-off level should be established for each objective." This last suggestion, if implemented, would counteract the type of confounding that leads to rectangular distributions and consequent difficulty in setting a cut-off level.

From this discussion it appears that generalizable rules for setting cut-off scores do not exist. Training developers setting cut-off scores must consider abilities of the trainee population, the complexity of skills and performances required by the objectives, through-put requirements of the training system, the minimum competence requirements, as well as a variety of other variables, and act accordingly.

Uses of CRM in Non-Military Education Systems

Prager et al. (1972) describe research on one of the first CRM systems planned for widespread implementation. This Individual Achievement Monitoring System (IAMS) was designed for the handicapped. Prager et al. point out that

standardized tests often are useless for handicapped individuals, since they have little value in directing remediation. Tests built to reflect specific objectives are much more useful when dealing with such populations. The use of CRM allows a handicapped child's progress to be related to criterion tasks and competency levels. CRM is further indicated by the need for individualized instruction and individualized testing when dealing with individuals who exhibit a variety of perceptual and motor deficiencies.

As a result, a CRM-centered accountability system was devised. This project began with the construction of a bank of objectives and test items to mesh with the type of diagnostic individualization required for education of the mentally handicapped. To meet these needs, the objectives were, of necessity, highly specified. The CRT-guided instructional system was geared to yield information to support three types of decisions: placement, immediate achievement, and retention. Standardized diagnostic and achievement tests were also used to aid in placement decisions. It is still too early to comment on the ultimate usefulness of this system.

More recently, Popham (1973) presented data on the use of teacher performance tests. These tests require a teacher to develop a "mini-lesson" from an explicit instructional objective. After planning the lesson, the teacher instructs a small group of students. At the conclusion of the "mini-lesson", students are given a post-test. Affective information is obtained by asking students to rate the interest value of the lesson. Popham suggested three potential applications of the teacher performance test:

1. A focusing mechanism. To provide a mechanism to focus the teachers' attention on the effects of instruction, not on "gee-whiz" methods.

2. A setting for testing the value of instructional tactics. The teacher performance test can be used as a "test bed" to evaluate the differential effectiveness of various instructional techniques. An important aspect of this application involves a post-lesson analysis in which the instructional approach is appraised in terms of its affects on learners.
3. A formative or summative evaluation device. Popham views this application of teacher performance tests as extremely important, particularly in the appraisal of in-service and pre-service teacher education programs.

Popham presented three applications of the teacher performance tests. The applications were for the most part viewed as effective, however a number of problems were revealed that may be symptomatic of performance tests in general. Popham found that unless skilled supervisors were used to conduct the mini-lesson, many advantages of the post-lesson analysis were lost. Popham also found that visible dividends were gained by using supplemental normative information to give teachers and evaluators additional information regarding the adequacy of performance.

In a similar area, Baker (1973) reported using a teacher performance test as a dependent measure in the evaluation of instructional techniques. Baker discussed shortcomings in using CRTs as dependent variables. These shortcomings are largely based on the peculiar psychometric properties of CRTs. However, Baker feels that CRM is valuable for research purposes, even with the large number of unanswered questions concerning their reliability and validity. Baker points out ". . . if the tests have imperfect reliability coefficients in light of imperfect methodology, the researcher is compelled

to report the data, qualify one's conclusions, and encourage replication." Baker also feels that the use of teacher performance tests with indeterminate psychometric characteristics is not ethically permissible for evaluation of individuals--at least for the present.

Millman (1973) described three studies on the psychometric characteristics of teacher effectiveness performance tests, using materials similar (mini-lessons) to those of Baker (1973) and Popham (1973). According to Millman, the most disturbing findings resulting from the studies were "the erratic and low test-retest reliabilities." Millman discussed several possible reasons for the discouraging reliability findings, but none of these seemed to ameliorate their significance, so he concluded that "clearly more definitive work is needed on teaching performance tests."

In a slightly different area of application, Knipe (1973) summarized the Grand Forks Learning System in which CRTs played a salient part. The Grand Forks School District began by creating detailed specifications of performance objectives in K-12 for most subject areas. These objectives were designed to form the basis of a comprehensive set of teacher/learner contracts, as one instructional method. It was found that mathematics was the subject area most amenable to analysis, and therefore it received the most extensive treatment. The mathematics test consisted of approximately 120 criterion-keyed items for each grade level 3-9. After extensive tryout, items were revised on the basis of teacher and student recommendations, as well as on the basis of a psychometric analysis. Knipe found that, teachers regarded the CRTs as useful in supplementing NRTs, and, in addition, found them useful for placement. Knipe concluded: "The criterion-referenced test is the only

type of test that a school district can use to determine if it is working toward its curriculum goals."

Klein and Kosecoff (1973) summarized present efforts in non-military CRM, emphasizing CRTs for mathematics. They described nine different CRTs, analyzing each as to their characteristics on five continua: program focus, instructional dependence, objective and item generation, test models and packaging, and test scores. The following CRTs and CRM programs were described: (1) "Prescriptive Mathematics Inventory", used to assess achievement on objectives associated with fourth-through-eighth grade mathematics curricula; (2) "Comprehensive Achievement Monitoring", a computer-assisted multipurpose evaluation system; (3) "Individualized Criterion Referenced Testing", currently available in kit form for assessing reading and mathematics skills for grades one through eight; (4) "Instructional Objectives Exchange", providing manuals covering objectives, sets of CRTs, and test guides; (5) "MINNEMAST Curriculum Project", CRTs designed to assess the MINNEMAST program, "a coordinated and sequential mathematics and science curriculum for the elementary school"; (6) "National Assessment of Educational Progress", CRTs designed to assess student achievement nationally, and available in forms for ages 9, 13, 17, and adult; (7) "Southwest Regional Laboratory", CRTs designed for quality assurance purposes in the development of text-referenced instructional management systems; (8) "System for Objectives Based Assessment--Reading", CRT items keyed to a set of performance objectives, and covering K-12 reading; and (9) "Zweig and Associates", CRTs indexed to prescription r teaching alternatives, and available for K-8 mathematics assessment.

Boyd and Shimberg (1971) developed a "Handbook of Performance Testing"

the majority of which is devoted to a portfolio of over 100 pages presenting a great variety of criterion-referenced performance tests. These tests, ranging from woodwork and metal repairs through dental hygiene to cosmetology, are presented in considerable detail, and are illustrative of creative approaches to the design of performance test items.

Hambleton (1974) has commented on CRM as the method of choice in evaluating individualized instruction programs. He has considered several such programs thoroughly, and has recommended three types of CR testing as appropriate: unit pretesting, unit posttesting and curriculum-embedded testing. Curriculum-embedded testing is the least important of the three, since decisions made on the basis of such tests affect the student for only a short period of time and there exists an additional check of mastery on the posttest. Unit pre and posttests are of concern for assigning students to instructional units and for assessing mastery. False positive errors on such tests are considered more critical than false negative errors by the author.

Sherman, Zieky and Fremer (1974) have reviewed the process of developing CRTs in the language area. Guidelines for task analysis are also presented. The work is a prodigious volume which discusses many aspects of CRT development in general terms, however the areas of fidelity and practical constraints surrounding performance item development are ignored.

Military Uses

Extensive experience with use of CRM was reported by Taylor, Michaels, and Brennan (1973) in connection with the Experimental Volunteer Army Training Program (EVATP). To standardize EVATP instruction, reviews, and testing, performance tests covering a wide variety of content were developed and

distributed to instructors. The tests were revised as experience accumulated; some tests were revised as many as three times. Drill sergeants used the tests for review or remediation, while testing personnel used them for general subjects, comprehensive performance, and MOS tests. The tests also provided the basis for the EVATP Quality Control System, which was intended to check on skill acquisition and maintenance during training.

Unfortunately, problems were encountered with the change in role required of the instructors and drill sergeants under the system of skill performance instruction and training. Considerable effort was required to bring about the desired changes in instructor role. The CRT-based quality control system performed its function well by giving an early indication of problems in the new instructional system. Evaluation of the performance-based system revealed clear-cut superiority over the conventional instructional system. The problems with institutional change encountered by these workers should be noted by anyone proposing drastic innovation where a traditional instructional system is well-established.

Pieper, Catrow, Swezey, and Smith (1973) presented a description of a performance test devised to evaluate the effectiveness of an experimental training course. The course was individualized, featuring an automated apprenticeship instructional approach. Test item development for the course performance test was based on an extensive task analysis. The task analysis included gathering many photographs of job incumbents performing various tasks. These photos served as stimulus materials for the tests and were accompanied by questions requiring "What would I do" responses, or identification of correct vs. incorrect task performance. All items were developed for audio-

visual presentation, permitting a high degree of control over testing conditions. Items were selected which discriminated along several criteria. Internal consistency reliability was also obtained.

A somewhat similar development project, entitled Learner-Centered Instruction (LCI), (Pieper and Swezey, 1972), also describes a CRT development process. Here, a major effort was devoted to using alternate form CRTs, not only for training evaluation, but also for a field follow-up performance evaluation after trainees had been working in field assignments for six months.

Air Force Pamphlet 50-58, the Handbook for Designers of Instructional Systems, is a five-volume document which includes a volume dealing with objectives and CRM. A job performance orientation to CRM is advocated. Specific guidelines for task analysis and for translating criterion objectives into test items are presented for both "hands-on performance" and written contexts. The document is a good guide to basic "do's" and "don'ts" in CRT construction.

A similar Army document, TRADOC Reg 350-100-1 presents guidelines for developing evaluation materials, and for quality control of training. The term is used interchangeably with "performance tests" and with "achievement tests" in this document. The areas of CRM, in particular, and of evaluation, in general, are given minimal coverage. CON Pam 350-11 is essentially a revision of TRADOC Reg 350-100-1, designed to be compatible with unit training requirements. This document, although briefly mentioning testing and quality control, presents virtually no discussion of CRM.

Various Army schools have developed manuals and guides for their own use in the area of systems engineering of training. The Army Infantry school

at Fort Benning, Georgia, for example, has published a series of Training Management Digests as well as a Training Handbook and an Instructor's Handbook. There also exist generalized guidelines for developing performance-oriented test items, in terms of memoranda to MOS test item writers and via the contents of the TEC II program (Training Extension Course). The Field Artillery school at Fort Sill, Oklahoma provides an Instructional Systems Development Course pamphlet as well as booklets on Preparation of Written Achievement Examinations and an Examination Policy and Procedures Guide in the gunnery department. The Armor school at Fort Knox, Kentucky, publishes an Operational Policies and Procedures guide to the systems engineering of training courses. Generally these documents provide a cursory coverage of CRT development.

The Army Wide Training Support group of the Air Defense school at Fort Bliss, Texas provides an interesting concept in evaluation of correspondence course development. Although correspondence course examinations are necessarily paper-and-pencil (albeit criterion-referenced to the extent possible), many such courses contain an OJT supplement which is evaluated via a performance test administered by a competent monitor in the field where the correspondent is working. This is a laudable attempt to move toward performance testing in correspondence course evaluation. A supplement to TRADOC Reg 350-100-1 on developing evaluation instruments has also been prepared. This guide provides examples of development of evaluation instruments in radar checkout and maintenance and in leadership areas.

A course entitled "Objectives for Instructional Programs" (Insgroup, 1972) which is used at a number of Army installations has provided a diagrammatic guide to the development of instructional programs. CRM is not

covered specifically in this document, nor is it addressed in the recent Army "state-of-the-art" report on instructional technology, Branson et al. (1973). However, a CISTRAIN (Coordinated Instructional Systems Training) course (Deterline and Lenn, 1972a, b), which is also used at Army installations for training instructional systems developers, does deal with CRT development and, in fact, provides instructions for writing items and for developing CRTs. The study guide (1972b) deals with topics such as developing criteria, identifying objectives, selecting objectives via task analysis, developing baseline CRT items, revising first draft items, and preparing feedback. This document provides a good discussion of CRT development in an overview fashion.

Swezey and Pearlstein's (1974) document, Developing Criterion-Referenced Tests, was prepared under contract to the Army Research Institute for the Behavioral and Social Sciences, and provides comprehensive descriptions of a process for the development, validation, and use of CRTs in military applications. The manual covers distinctions between CRM and NRM, applications of CRTs, assessing adequacy of objectives, development of thorough test plans, construction of item pools, selection of "best" items by item analysis and and item review procedures, administration and scoring of CRTs, and assessment of CRT reliability and validity. The procedures for CRT development presented therein were derived from a comprehensive review of CRM literature,

U.S. Army FM 21-6 has recently undergone a comprehensive revision to suit the needs of field trainers. The revised manual is generally in tune with contemporary training emphasis, with considerable information on individualized training and team training. In particular, the extensive guidance

provided on generation of objectives should prove very useful to field trainers. While the revised FM 21-6 does not specifically refer to CRM, the obvious emphasis on NRT, which distinguished its earlier version, is gone. A possible weakness in the revised version is the tacit assumption that all trainees will reach the specified standard of performance. Although the requirement that all trainees reach criterion is not by itself unreasonable, practical constraints of time and cost sometimes dictate modified standards (e.g., 80% reaching criterion), just as Board actions or career reassignment may also affect the percentage of trainees reaching criterion. Where it is not feasible to wash-out or to recycle trainees, then remediation must be designed to permit an economical solution. FM 21-6 does not seem to address the remediation problem. In general, though, FM 21-6 is a good working guide to field training. It will be interesting to see how effective it is in the hands of typical field training personnel.

The use of CRTs in military operations has been slowed by the high initial cost of developing criterion-referenced performance tests. Often the use of CRTs for performance assessment has required operational equipment or interactive simulators, drastically raising costs. A solution to the cost problem may be found in the notion of Osborn (1970) who has devised an approach to "synthetic performance tests" which may lead to lowered testing costs, although little concrete evidence has appeared in the literature to date.

From these limited examples, it appears that the civilian sector has led in the research of methodological and theoretical questions concerning the use of CRM. However, the military has clearly led in the development and practical application of CRM.

Indirect Approach To CRM

Fremer (1972) suggested that it is meaningful to relate performance on Survey Achievement tests to significant real-life criteria, such as minimal competency, in a basic skills area. Fremer discussed various ways of relating survey test scores to criterion performance. All of these approaches are aimed at criterion-referenced interpretation of test scores. Fremer proposed that direct criterion-referenced inferences about an examinee's abilities need not be restricted to tests that are composed of actual samples of the behavior of interest. He suggested that considerable use can be made of the relationships observed among apparently diverse tasks within global content areas. Fremer argued further that tasks which are not samples of an objective may provide an adequate basis for generalization to that objective. He noted that, given a nearly infinite population of objectives, the use of a survey instrument as a basis for making criterion-referenced inferences would allow increased efficiency.

An example was presented, using a survey reading test to make inferences about ability to read a newspaper editorial. A CRT of ability to read editorials might consist of items quite different from the behavior of interest. Fremer offered the illustrative example, of using vocabulary test scores to define objective-referenced statements of ability to read editorials. He noted, however, that the usefulness of interpretive tables, i.e., those that provide statements referencing criterion behaviors to a range of test scores, depends heavily upon the method used to establish the relationship between the survey test scores and the objective-referenced ability. An essential aspect would be the use of a large and broad enough sample of

criterion performance to permit generalization to the broader range of performances.

Fremer's example provided for the definition of several levels of mastery and pointed out that an absolute dichotomy, mastery versus non-mastery, will seldom be meaningful. It is difficult to understand why Fremer made this statement, as the basic use of CRT is to decide whether an individual possesses sufficient ability to be released into the field or requires further instruction. Many levels of performance can be identified, but are ultimately reduced to pass-fail or to mastery/non-mastery. Fremer apparently based his objection on measurement error which can render classification uncertain. However, as discussed earlier, proper choice of cut-off and careful attention to development should minimize classification errors. Further, classification according to levels in addition to mastery/non-mastery would only increase the probability of classification errors. Fremer also proposed that the notion of minimal competency should encompass a variety of behaviors of varying importance, and that the metric of importance will vary with the goals of the educational system.

Fremer (1972) also set forth a method for relating survey test performance to minimal competency standards that involves a review of the proportion of students who are rated as failures at some point in the curriculum. This serves as a rough estimate of the proportion of students failing to achieve minimal competency. It is then possible to apply this proportion to the score distribution for the appropriate test in a survey achievement test, clearly a normative approach.

A second approach to referencing survey achievement tests to a criterion of minimal competency is to acquire instructor judgment about the extent to which individual items could be answered by students performing at a minimal level. By summing across items, it is possible to obtain an estimate of the expected minimum score. Fremer however, recognized the limitations of this latter process with its high reliance on informed judgment. A third method proposed by Fremer, seeks to define minimal competency in terms of student behaviors. The outcome of this method is the identification of bands of test scores associated with minimal competency. As in the second method, processes involved in this method also rely on informed judgment.

Still another method proposed by Fremer involves developing new tests with a very narrow focus, i.e., a smaller area of content and a restricted range of difficulty. Using this method, it is not necessary to address every possible objective, however, a test composed of critical items can be developed by sampling from the item pool. The next step in the process involves relating achievement at various curriculum placements to the focused test and the survey instrument. This allows keying of the items on the survey test to specific critical objectives. A final method put forth by Fremer is the stand-alone work sample test. This technique is intended for use when there is an objective that is of such interest that it should be measured directly.

The procedures ennunciated by Fremer are clever in concept, but are mainly applicable to school systems, and traditional curricula, where well-developed survey instruments exist. Even where appropriate survey instruments exist, considerable work is involved in keying the survey instrument.

In non-school system instructional environments, dealing with non-traditional curricula, it is unlikely that appropriate survey instruments exist.

Gray (1973) developed a written CRT designed to assess performance on the Piagetian tasks of pendulum oscillation, equilibrium in the balance, and combinations of colorless and colored chemical bodies. Ninety-six subjects--12 in each of 8 age groups (9-16 years old)--were administered the written test and the actual Piagetian tasks in a counterbalanced design. Gray's statistical analysis revealed that, in most cases "the correspondence between the predicted and written item sequences is excellent." He concluded that "the correlations between the two methods measuring the same set of developmental logic (validity values) along with moderate reliabilities are encouraging. . . [and] support the conclusion that a written test using the developmental logic postulated by Piaget as its behavioral criterion is definitely possible. . ." Although Gray noted that there was considerable room for improvement in this particular attempt at test development, the implications for an indirect approach to CRM are obvious.

Using NRTs To Derive CRM Data

Cox and Sterrett (1970) proposed an interesting method for using NRTs to provide CRM information. The first step in this method is to specify curriculum objectives and to define student achievement with reference to these objectives. The second step involves coding each standardized test item with reference to curriculum objectives. With coded test items and knowledge of the position of each student in the curriculum, it is then possible to determine the item validity, in the sense that students should be able to correctly answer items that are coded to objectives that have

already been covered. Step three is scoring the test independently for each student, taking into account position in the curriculum. The authors suggested that this model is particularly applicable to group instruction, since placement in the curriculum can generally be regarded as uniform. Therefore, it is possible to assign each student a score on items whose objectives he has covered. It is also possible to obtain information on objectives which were excluded or not yet covered.

Livingston (1972c) delineated a method for computing criterion-referenced indices from a set of norm-referenced test scores. First, the norm-referenced mean (μ_x), variance [$\sigma^2(X)$], and reliability coefficient ($\rho^2(X,T_x)$] are computed. Then, formulae are used for conversion to criterion-referenced indices. For example, the criterion-referenced reliability coefficient [$k^2(X,T_x)$] is found by the following formula:

$$k^2(X,T_x) = \frac{\rho^2(X,T_x)\sigma^2(X) + (\mu_x - C_x)^2}{\sigma^2(X) + (\mu_x - C_x)^2}$$

where: C_x = the criterion score

The appropriateness of Livingston's techniques have yet to be empirically verified, however.

Considerations for a CRT Implementation Model

The development and use of CRM is a fairly recent occurrence in instructional technology. Partially as a result of this, there is no comprehensive theory of CRM, such as exists for NRM. Hence, the concepts of CRT validity and reliability are not yet well developed.

The need for content validity in CRT is well recognized, however. But there is no single CRT construction methodology which will serve for all content domains. Unresolved questions also exist in the area of Bandwidth fidelity, and the use of reduced fidelity in criterion-referenced performance tests.

The rationale for the use of CRM in evaluating training programs and describing individual performance is well established. For example, the instructional systems development model developed by Branson, Hannum, Rayner, and Johnson (1974), and intended for implementation throughout the Armed Services, uses CRM as an integral part. Branson et al. noted that "The process involved in the development of objective-referenced tests is the development of test exercises that measure student performance of a specific element identified in the analysis of the learning requirements. . .," and that "the test exercises and learning objectives must be in agreement and must reflect the specific learning elements that were identified in the learning analysis step [of the instructional systems development model]."

To ensure the best possible results, military or industrial users should exert every effort to maintain stringent quality control, including:

1. Careful task analysis:
    a. Observation of actual job performance when possible
    b. Identification of all skills and knowledges that must be trained
    c. Careful identification of job conditions
    d. Careful identification of job standards
    e. Identification of critical tasks.

2. Careful formulation of objectives

   a. Particular care in the setting of standards

   b. Accurate identification of all objectives

   c. Independent checks on the content of the objectives

   d. Special attention to critical tasks.

3. Item development

   a. Determine if all objectives must be tested

   b. Survey of resources for test

   c. Development of item sampling strategies

   d. Determination of appropriate item format

   e. Development of item pool for objectives to be tested

   f. Development of a tryout plan and criteria for item acceptance

   g. Tryout of items

   h. Revision or rejection of unacceptable items.

4. Consideration of reliability and validity

Particular care must be exercised in setting item acceptance criteria for item tryout. The use of typical NRT item statistics should be minimized. Many usual methods are not adequate, e.g. internal consistency estimates. Traditional stability indexes may also be inappropriate, due to small numbers of items and reducted variance.

By adhering to strict quality control measures, it should be possible to obtain measures that have a strong connection with a specified content domain. Whether they are sensitive to instruction, or will vary greatly due to measurement error, is unknown. Careful tryout and field follow-up may currently be the best controls over errors of misclassification due to

poor measurement. The ethical question of the use of measures with unknown psychometric properties in making decisions about individuals remains to be addressed.

Cost-Benefits Considerations

Although the costs of training and the costs of test administration can readily be quantified in dollar terms, we lack an adequate metric to rigorously assess costs of misclassification. Emrick (1971) proposed a ratio of regret to quantify relative decision error costs. Emrick's metric however, appears rather arbitrary and in need of further elaboration. The probability of misclassification is the criterion against which an evaluation technique must be weighed. The results of misclassification range from system-related effects to interpersonal problems. In some instances where misclassification results in a system failure, cost can be accurately measured, and is likely to be high.

A relative index of cost can be gained from task analysis. If the analysis of the job reveals large numbers of critical tasks, or individual tasks whose criticality is very high, then the cost of supplying a non-master can be assessed as high, and great effort is justified in developing high fidelity CRTs in conjunction with a training program. Misclassification also results in job dissatisfaction and morale problems, evidenced by various symptoms of organizational illness, e.g., absenteeism, high turnover, poor work group cohesion, etc.

A possible solution to the cost-benefit dilemma may come from work with symbolic performance tests and the work cited earlier showing that job knowledge tests can sometimes suffice. The use of symbolic tests and/or job

knowledge tests may result in reduced testing costs in some instances. However, development of suitable symbolic performance tests may prove to be difficult. And, as progress is made in lowering CRM development cost, cost-benefit problems will be largely obviated.

As the question currently stands, there is no doubt that CRM provides a good basis for evaluation of training and the determination of what a trainee can actually do. If the system in which the trainee must function produces a number of critical functions which will render misclassification expensive, then CRM is a must. And, if the system has been developed from task analytic data, CRM development is both desirable, for evaluation purposes, and cost-effective, whether or not there are many critical tasks involved.

Brief Summary of the State-of-the-Art in Criterion-Referenced Testing

Now, let us set forth a general position on theoretical and technical aspects of CRT construction and use, based upon the state-of-the-art of CR testing as we see it. Positions are presented sequentially for the following topics:

1. Design considerations and CRT use
2. Construction methodology and related issues
3. CRT administration and scoring
4. Reliability and validity

Design Considerations and CRT Use

Among the major considerations in CRT construction is the way in which specific uses may affect test design. Test design may vary in several related fundamental respects, such as the basis upon which test items are constructed and selected. In CR testing, items are generally developed from an analysis of tasks to be performed and from attempts to operationally define the behaviors required. This is not necessarily the case in norm referenced (NR) testing. The manner in which scores are interpreted and used also differentiates CRTs from NRTs. In CR testing, scores attained by examinees are interpreted against an external, absolute standard—as opposed to the distribution of scores attained by other examinees; which is the case with NRTs.

It must first be decided whether a CRT, as opposed to a NRT, is appropriate. CRT scores do not lend themselves to ordering individuals along a continuum, thus if the primary use of test results is to select among individuals for promotion, special honors, etc., CR testing is contraindicated. Whenever information is desired for purposes of comparing examinees, NR testing appears to be more appropriate than CR testing. This applies to tests of achievement, knowledge, and performance.

CR testing is usually the technique of choice when evaluations are to be made on the basis of an individual's achievement of specific objectives. Here the primary question of interest is: "How well can an individual perform relative to an external standard?", rather than: "How well does an individual do compared to others?".

Cost Effectiveness

CRTs may be more expensive to develop and administer than NRTs, in terms of absolute costs. CRT-specific development costs are due largely to the need for carefully deriving and specifying objectives, while additional administration costs may result from the necessity of comparing examinee performance to external standards. Nevertheless, CR testing may well be more cost-effective in the long run, if there is a genuine need to ascertain an individual's ability to perform a specific task.

Indirect approaches to criterion-referencing, by correlating symbolic performance and/or job knowledge test results with performance measures, may be an approach to alleviating the high costs of CRTs. Such approaches involve the development of two tests at different levels of fidelity for each objective, and subsequent validation of the indirect measures against the performance measures. Justification for these approaches center on savings in administration time and costs.

Development of direct CRTs appears justified, desirable and cost-effective, if there is a need to ensure that individuals will be able to perform adequately on the tasks for which they are being trained. When there is a need for ensuring minimal, absolute levels of performance, CR testing is the approach of choice.

Screening and Diagnosis

CRTs are applicable for use as screening devices in cases where there is a possibility that individuals may be able to perform tasks without training. If a person can achieve the criterion level on a CRT, he should be able to enter the job without intervening training. Similarly, CRTs may be used to determine the appropriate point i a training cycle for an individual to commence training.

CRTs may also be used as diagnostic aids. Persons achieving the criterion level might be channeled into advanced instruction, or remediation might be suggested for those falling below criterion level on certain objectives. CR testing for diagnostic purposes is likely to be more difficult and more expensive than CR testing for achievement of objectives, because detailed documentation on the examinees' behavior is required. This may necessitate more examiners and/or more elaborate schemes for collecting data.

Evaluation of Instructional Programs

Aside from the assessment of individual performance against absolute standards, CRTs may also be used to evaluate instructional programs. Here, the primary question of interest is: "Has my instructional program taught what it is supposed to teach?". NR testing is less appropriate for such an application than is CR testing, since wide score ranges before and after administration of the instructional program are not necessarily germane to the question of interest. CRTs designed for this application are presumably based directly upon instructional objectives since the basic question is whether or not the program has successfully taught performance compatible with the instructional objectives. CRTs thus provide data having direct relevance to the question.

Construction Methodology and Related Issues

Due to the relative recency of the CR testing concept, many theoretical and practical aspects of CRT construction methodology are not so well defined as is the case for NRTs. Additional sophistication in CRT construction methodology must await further research on theoretical issues, and results from more extensive attempts at CRT implementation. Nevertheless, some general "do's and don'ts" for CRT construction can be extracted from the methodological literature.

Task Analysis

First, CRT construction requires careful analysis of the tasks comprising the test's subject. While conduct of the task analysis itself may be outside the test developer's domain, the test developer must obtain analytic data on: (1) skills and knowledges necessary for task performance, (2) required performances stated in behavioral terms, (3) criteria associated with each identified performance, and (4) conditions under which the tasks must be performed.

Without these data, the test developer cannot adequately define objectives, and consequently cannot match test items to objectives. Nor can he ensure the content validity of the test. If usable CRTs are to be constructed, task analyses are necessary prerequisites.

Preparing Objectives

Preparing objectives is one of the first formal steps in constructing a CRT. Mager (1962) has documented a useful procedure for formatting these objectives. Mager's suggestions for structuring objectives also appear

appropriate. Information to be used in preparing objectives is best derived from thorough task analytic data.

If the test developer's input includes a list of unitary objectives--objectives covering separate, single tasks--the test developer's primary task is to match test items to these objectives. The test developer must assume that objectives are properly matched to the actual job tasks. If this assumption is violated, the resulting CRT will lack content validity. If however, the assumption is accurate, and the developer properly matches items to objectives, content validity will be achieved. Thus, the test developer must be knowledgeable about appropriate formats and quality standards for objectives in order to make an adequate assessment of their suitability for CRT development.

Matching Items to Objectives

Mager (1973) has provided a sound plan for matching CRT items to objectives. Mager's plan involves matching performances and conditions stated in, or implied by objectives, with corresponding item performances and conditions. Mager's plan omits a procedure for matching standards among objectives and test items, however implies that standards should also be matched.

The test constructor's task is to create test items that are congruent with objectives. To the extent that objectives are "fuzzy", the test constructor cannot create appropriate items. It is recommended that he send fuzzy objectives back to their originator, annotating their difficulties and requesting a reconsideration.

When the test developer has received an adequate objective (or set of objectives) for which a test is to be constructed, a number of factors must be considered before items are matched to objectives. These factors include: practical constraints in the testing situation, test fidelity, test format, and number of items required to test a given objective.

Practical constraints must be systematically assessed before test items can be constructed so that the items can be built with performance indicators which are suitable for such considerations as: testing conditions, tester availability, time availability, facility and equipment availability, etc. These considerations obviously impact on test fidelity. CRT items should be constructed at the highest level of fidelity practicable, consistent with situational constraints. In cases where critical objectives are to be tested, special care must be taken to develop sufficiently high fidelity items so that critical task mastery can be accurately assessed.

Selecting Among Objectives

The tactic of selecting among objectives, that is, randomly testing a subset of objectives, may be used in some instances, as long as trainees do not know the subset to be tested. This tactic must not be used when critical objectives are involved. For objectives of a non-critical nature, selection may be used to overcome practical constraints imposed by the testing situation, without necessitating modification of objectives. Selection among objectives should never be done when it is necessary to certify that individuals qualify on all objectives.

Number of Items

No hard and fast rules for specifying the number of items to be created for a given objective exist. It is recommended that as many items as test situation time availability will permit, within limits suggested by considerations of motivational and fatigue factors, should be included. As Graham (1974) has noted, "even for highly homogeneous tests, four or five items may be necessary to minimize classification errors." Thus, even for CRTs measuring a single, well-specified objective with few confounding factors, additional items may help to reduce measurement error. For more heterogeneous tests, the desireability of having extra items may be even more pronounced.

Format

Test format may, in many cases, be largely dictated by objectives. Certain objectives for example, may require hands-on performance testing. Such things as number of items to be included, and practical constraints such as time and manpower availability, may also help determine format—e.g., a situational item, multiple-choice format might be the only feasible way of testing some sets of objectives. A general guideline might be based on Edgerton's (1974) suggestion, that item styles not be mixed in the same test, so as to avoid measuring "test taking skill" instead of subject matter competence.

Item generation rules, such as "item forms" and "facets" are not yet sufficiently researched to warrant use by personnel who are not sophisticated in psychometrics. Hence, for objectives that may be tested by

an unlimited number of items, such as those dealing with concepts, the best suggestion that can be offered testing personnel at this time, is to be sure that each item matches the objective it tests.

Item Pools

After the test developer has considered such factors as fidelity, number of items, etc., items can be matched to objectives using principles similar to those advanced by Mager (1973). The test developer should construct a pool of items considerably larger than the number required for the test, so that the best items can be selected. Items are then constructed at the level of fidelity and in the format previously determined.

Item Analysis

Traditional item analysis techniques, like other statistical techniques developed in conjunction with NR testing, have limited applicability for CR testing (due to restricted ranges of score variance in CRTs). Although recent studies have suggested techniques for increasing variance of CRT scores (e.g., Haladyna, 1973; Woodson, 1973) these techniques are "experimental", and it is not yet appropriate to apply them as a matter of course. Consequently, until additional research develops and refines new approaches to item analysis appropriate for CR testing, a simple index which relies on the use of "masters" and "non-masters" (e.g., those who are beginning training and those who have completed training) appears to be an appropriate technique.

"Masters" and "non-masters" are tested and their patterns of pass and fail on the items are recorded. ϕ coefficients are computed using

four-fold tables ("master"-"nonmaster", pass-fail) for each item. Good items are those which are passed by "masters" and failed by "nonmasters." Items are poor if there is little difference on pass-fail patterns between "masters" and "nonmasters", or if more "nonmasters" than "masters" pass them. Low or negative ϕ coefficients act as warning flags. Items receiving low coefficients should either be thrown out or, at least, reconsidered carefully before inclusion in a CRT. These warning flags are relevant if the pool of items is homogeneous, or if it is composed of items testing several objectives.

Care must be exercised to ensure that all objectives are represented by the proper number of items, as determined previously. Item balance among disparate objectives measured by the same test should be maintained as planned.

CRT Administration and Scoring

Administration

Like all tests, CRTs must be administered under standardized conditions. CRTs should include accompanying documentation which specifies: (1) test administration conditions; (2) instructions; (3) administration procedures (including how to handle questions, how to check and set up test supplies and equipment, etc.; (4) circumstances for excusing examinees from the test, due to illness, fatigue, etc.; (5) environmental circumstances under which test administration should be cancelled; and (6) scoring procedures.

Test administrators must be trained to follow specifications precisely. Since specifications will apply to any test, documentation accompanying a specific CRT need not necessarily be extremely detailed—except for special requirements such as setting up the test facility, and test scoring.

Scoring

Test scoring procedures must be developed during the test construction process, since they will generally vary as a function of the type of CRT. There are a number of interrelated decisions that must be made concerning scoring. These include:

1. Objectivity of scoring
2. Process vs product scoring methods
3. Type of scoring (go/no-go, rating scales, etc.)
4. Cut-off points
5. Non-interference vs assist methods

Objectivity

Every attempt should be made to maximize objectivity in scoring CRTs. In low fidelity tests, such as those using multiple-choice formats, objectivity is apparent. (Such tests can be computer-scored.) In higher fidelity CRTs, it is relatively simple to maximize objectivity for hard-skill subjects, however soft-skill areas, such as tactics, leadership, etc. are more difficult to test objectively. To the extent that objectivity is not achieved, reliability is attenuated. Efforts must be made to specify soft-skill objectives precisely, so that appropriate items (with associated objective scoring procedures) can be prepared. Even in the best of circumstances, however, soft-skill CRTs will probably have less objective scoring guides than will tests of hard-skill subjects. One way to maximize objectivity in soft-skill CR testing is to require several raters to assess each individual. Inter-rater reliabiltiy can then be calculated. If low inter-rater reliability is found consistently, the test should be revised.

Process-Product

R.C. Smith's (1965) guidelines for determining process versus product measurement appear adequate, with slight modifications. That is, product measurement is always appropriate if the objective specifies a product. When a product measure is called for, it should be incorporated into the objective, and carried over into the test items. Product measures are called for when:

(a) the product can be measured as to presence or characteristics

(b) the procedure leading to the product can vary without affecting the product.

Process measurement is indicated when the objective specifies a required sequence of performances which can be observed, and the performance is as important as the product. Process measurement is also appropriate in cases where the product cannot be measured for safety or other constraining reasons.

There may also be situations where both process and product measurement are appropriate for a given objective. Following are several examples of conditions that may call for both product and process measurement:

(a) Although the product is more important than the process(es) which lead to its completion, there are critical steps which, if misperformed, may cause damage to equipment or injury to personnel.

(b) The process and product are of similar importance, but it cannot be assumed that the product will meet criterion levels.

(c) Diagnostic information is needed. (By having process as well as product measures, information as to why the product does not meet the criterion can be obtained.)

When both process and product measures are obtained for a specific objective, scoring must follow the criterion specified by the objective. That is, if the criterion specifies only a product, then process scores should not be used to assess achievement of the criterion.

Type of Scoring

The type of scoring system employed must be appropriate for the objective. If the objective specifies an action or product, a go/no-go scoring system should be used (either the action occurs in the proper sequence or it does not; either the product results or it does not). If the objective specifies characteristics of a criterion-level product or action, a rating scale or other form of point assignment is indicated. Point assignments must be made on an explicit, well-defined basis for each item. For rating scales, inter-rater reliability must be high. Point assignments must be tied to criterion levels specified in the objective.

Cut-Off Points

Cut-off levels should reflect mastery of the objective to the extent required. Since factors other than ability to perform a task (such as careless errors, measurement errors, etc.) may affect an individual's score, cut-off levels are often set somewhat below 100 percent. If, for example, an objective calls for multiplication of two four-digit numbers, the criterion might specify performing 10 such sets within five minutes,

achieving the correct answer in at least eight cases. Thus, the cut-off score of 8 (below 8 = fail) reflects an arbitrary definition of mastery. True mastery would require 10 out of 10.

Graham (1974) has made some valuable suggestions concerning the setting of cut-off points. The cut-off, basically, should discriminate masters from non-masters. However, as item domains become more broad, more heterogeneous item sets are required. Thus, the confounding influence of skills and knowledges which are not directly related to objectives increases. For tests measuring objectives having broad domains (or several objectives with different domains) the overlap between mastery and non-mastery scores consequently widens.

When little overlap occurs between mastery and non-mastery scores (as is the case for tests measuring a single objective with a relatively restricted domain) setting a cut-off score is less critical. The cut-off point should reflect the standard specified by the objective, and can do so without falling into the zone of overlap between masters and non-masters, since this zone, by definition, is either narrow or non-existent. On the other hand, if the overlap is wide, the point at which the cut-off score is set, is critical. Wherever the cut-off score is set, there will be some misclassification. In such cases, there are two considerations. First, objectives must be specified precisely, with item domains as restricted as possible, in order to narrow the mastery-nonmastery overlap. When achievement of several objectives of disparate nature are measured by a single test, separate scores for each objective's item set should be obtained, each with its own cut-off. However, for end-of-course or end-of-cycle

exams which assess high levels of skill and knowledge integration, a single cut-off may be set, since what is to be evaluated is a cluster of skills and knowledges applied in combination.

Second, costs of false positives and false negatives must be considered. If the costs for false negatives are relatively high (e.g., manpower needs are critical) the cut-off score might justifiably be lowered. If the costs of false positives are high, then cut-off scores must remain high. In any case, when performance on critical tasks is tested, cut-off points must be kept high enough to reflect the standards specified in the objectives for those tasks.

Assist vs Non-Interference

In general, a non-interference method of test administration is preferred over an assist method, in CR testing applications. In the assist method, the examinee is scored no-go for a missed item, corrected, and then allowed to proceed. A major problem here, is that if the criterion requires an examinee to complete a chain of steps, he should be tested on to his ability to do so. On the job, the examinee will have to complete the chain of steps correctly, with no help. There are however, cases in which an assist scoring technique can be profitably used. These involve uses of CR testing for diagnosis. In such cases, the trainee is permitted to complete a chain of steps and given assistance on those which he cannot perform adequately. He is typically scored no-go for steps where he is assisted. The record of no-go steps is a useful diagnostic tool—remediation

can concentrate on missed steps. Such records may also be useful for evaluating instructional material, especially if many examinees have similar patterns of no-go items.

Reliability and Validity

Reliability

Techniques for assessing CRT reliability are, for the most part, either not fully developed or are based on questionable assumptions. (For example, see Livingston, 1972; Oakland, 1972; Haladyna, 1974; and Woodson, 1974). The need for additional work in the area of CRT reliability, continues to be a pressing one.

A practical solution is to assess test-retest reliability of CRTs, a procedure which does not depend on internal consistency, and which increases the variability of test results, because of the two test administrations required. The ϕ coefficient is useful for analyzing the resulting four-fold (first administration-second administration, pass-fail) data. ϕ values less than +.50 would indicate unacceptable test-retest reliability for CRTs.

Validity

Content validation is an especially appropriate method in CRT applications. A CRT is content valid if the test items are carefully based on the performances, conditions, and standards specified in the objectives <u>and</u> if the test items appropriately sample objectives. (Of course, the objectives themselves must be sound.) Thus, in most instances, careful test construction will, itself, enable the development of content valid CRTs.

However, in instances where low fidelity CRTs are constructed, it may be more difficult to determine content validity, since the items are not likely to be precisely matched to objectives. In such cases, there are two additional types of criterion-related validation that are well-suited to CRTs: concurrent validity and predictive validity.

In determining concurrent validity, CRT results are compared with an outside measure of the behaviors tested by the CRT. This outside measure must be the best available assessment of performance on the objective(s) in question. The assessment of concurrent validity, involves individual assessment via the CRT and the outside measure close together in time (concurrently). ϕ again is used on the four-fold data (CRT-other measure, pass-fail).

Predictive validity involves the same assumptions. The outside measure must be an accurate measure of the performance in question, or the validation will be meaningless. Predictive validity is calculated the same way, except the outside measure is taken at a later time—i.e., when the individuals are actually performing the job for which they've been trained. The ϕ estimate is calculated just as for concurrent validity.

References

American Psychological Association. Standards for Educational and Psychological Tests. Washington, D.C.: American Psychological Association, 1974.

Ammerman, H.L. & Melching, W.H. The Derivation, Analysis and Classification of Instructional Objectives. HumRRO Technical Report 66-4, 1966.

Baker, E.L. Using measurement to improve instruction. Paper presented at the Annual Meeting of the American Psychological Association, Honolulu, 1972.

Baker, E.L. Teaching performance tests of dependent measures in instructional research. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, 1973.

Besel, R.R. Program for computing least squares estimates of item parameters for the mastery learning test models: Fixed GMP version. SWRL, 1973. (a)

Besel, R.R. Program for computing least squares estimates of item parameters for the mastery learning test models: Variable GMP version. SWRL, 1973. (b)

Block, J.H. (Ed.) Mastery Learning: Theory and Practice. New York: Holt, Rinehart, and Winston, 1971.

Boyd, J.L. Jr. and Shimberg, B. Handbook of Performance Testing: A Practical Guide for Test Makers. Princeton, New Jersey: Educational Testing Service, January, 1971.

Branson, R.K., Hannum, W.H., Rayner, G.T., and Johnson, B.F. The Instructional Systems Development Model: A Description of the Processes, Procedures and Products. Talahassee, Florida: Center for Educational Technology, Florida State University, March, 1974.

Branson, R.K., Stone, J.H., Hannum, W.H. & Rayner, G.T. Analysis and assessment of the state of the art in instructional technology. Final Report: Task 1 on Contract No. N61339-73-C-0150, U.S. Army Combat Arms Training Board and The Florida State University, 1973.

Carver, R.P. Special problems in measuring change with psychometric devices. Evaluative Research: Strategies and Methods. Washington: American Institute of Research, 1970.

Carver, R.P. Two dimensions of tests: Psychometric and edumetric. American Psychologist. July, 1974, 512-518.

Chenzoff, A.P. A Review of the Literature on Task Analysis Methods. Valencia, Pennsylvania: Applied Science Associates, Inc., Technical Report 1218-2, 1964.

Cox, R. C. & Sterrett, B. G. A model for increasing the meaning of standardized test scores. Journal of Educational Measurement, 1970, 2, 227-228.

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

Cronbach, L. J. Evaluation for course improvement. Teachers College Record, 1963, 64, 672-683.

Cronbach, L. J. The dependability of behavioral measurement: Theory of generalizability for scores and profiles. New York: Wiley, 1972.

Cronbach, L. J. & Gleser, G. C. Psychological tests and personnel decisions. Urbana: University of Illinois Press, 1965.

Cronbach, L. J. & Meehl, P. E. Construct validity in psychological tests. Psychological Bulletin, 1955, 92, 281-302.

Davies, I. K. Task analysis: Some process and content concerns. AV Communication Review, Spring 1973, 21, No. 1, 73.

Department of the Army. How to prepare and conduct military training. FM 21-6, Washington, D.C.: Headquarters, Department of the Army, 20 Jan. 1967.

Department of the Army. Systems engineering of unit training. CON Pam 350-11 Fort Monroe, Va.: Headquarters, U.S. Continental Army Command, 12 January 1973.

Department of the Army. How to prepare and conduct military training. FM 21-6, Washington, D.C.: Headquarters, Department of the Army, 1 Dec. 1973.

Deterline, W. A. & Lenn, P. D. Coordinated instructional systems: Lesson book. Palo Alto, Calif.: Sound Education, Inc., 1972a.

Deterline, W. A. & Lenn, P. D. Coordinated instructional systems: Study resource materials book. Palo Alto, Calif.: Sound Education, 1972b.

Ebel, R. L. Content standard test scores. Educational and Psychological Measurement, 1962, 22, 15-25.

Ebel, R. L. Criterion-referenced measurement: Limitations. School Review, 1971, 79, 282-288.

Edgerton, H. A. Personal communication, 1974.

Edmonston, L. P., Randall, R. S., & Oakland, T. D. A model for estimating the reliability and validity of criterion-referenced measures. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, 1972.

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

Engel, J. D. & Rehder, R. J. A Comparison of Correlated-Job and Work-Sample Measures for General Vehicle Repairmen. HumRRO Technical Report 70-16, 1970.

Examination policy and procedures guide. Fort Sill, Okla.: U.S. Army Field Artillery School Gunnery Department, 23 March 1973.

Flanagan, J. C. Critical requirements: A new approach to employee evaluation. Personnel Psychology, 1949, 2, 419-425.

Flanagan, J. C. Discussion of symposium: Standard scores for aptitude and achievement tests. Educational and Psychological Measurement, 1962, 22, 35-39.

Folley, J.D., Jr. The learning process. In R.L. Craig and L.R. Bittel (eds) Training and development handbook. New York: McGraw-Hill, 1967.

Frederiksen, N. Proficiency tests for training evaluation. In R. Glaser (Ed.) Training Research and Education. Pittsburgh: University of Pittsburgh Press, 1962.

Fremer, J. Criterion-Referenced Interpretation of Survey Achievement Tests. ETS Development Memorandum, TDM-72-1, 1972.

Fremer, J. & Anastasio, E. Computer-assisted item writing - I (spelling items). Journal of Educational Measurement, 1969, 6, 69-74.

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

Glaser, R. & Nitko, A. J. Measurement in learning and instruction. In R. L. Thorndike (Ed.) Educational Measurement. Washington: American Council on Education, 1971, 625-670.

Goodman, L. A. & Kruskal, W. H. Measures of association for cross classification. American Statistical Association Journal, 1954, 49, 732-764.

Graham, D.L. An examination of the feasibility of using criterion-referenced measurement in large-scale, survey testing situations. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, April 1974.

Gray, W.M. Development of a Piagetian-based written test: A criterion-referenced approach. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, February, 1973.

Guion, R.M. Open a new window: Validities and values in psychological measurement. American Psychologist. May, 1974, 287-296.

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

Haladyna, T.M. Effects of different samples on item and test characteristics of criterion-referenced tests. Journal of Educational Measurement. 1974, 11(2), 93-99.

Hambleton, R.K. Testing and decision-making procedures for selected individualized instructional programs. Review of educational research, 1974, 44(4), 371-400.

Hambleton, R.K. & Garth, W.P. Criterion-referenced testing: Issues and applications. Amherst: Massachusetts University, School of Education, 1971.

Hambleton, R. K. & Novick, M.R. Towards a theory of criterion-referenced tests. American College Testing Technical Report, Iowa City, 1971.

Hambleton, R. K., Rovinelli, R., & Gorth, W. P. Efficiency of various item-examinee sampling designs for estimating test parameters. In proceedings, 29th Annual Convention of American Psychological Association, 1971.

Handbook for designers of instructional systems. AF Pamphlet 50-58, Wright-Patterson Air Force Base, Ohio, 1973.

Hanson, R. A. & Berger, R. J. Quality assurance in large scale installation of criterion-referenced instructional programs. Paper presented at the Annual Meeting of the National Council of Measurement in Education, New York. 1971.

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

Helmstadter, G. C. A comparison of traditional item analysis selection procedures with those recommended for tests designed to measure achievement following performance-oriented instruction. Paper presented at the Annual Meeting of the American Psychological Association, Honolulu, 1972.

Hively, W., Patterson, H. C., & Page, S. A universe-defined system of arithmetic achievement tests. Journal of Educational Measurement, 1968, 5, 225-290.

Hively, W. W. II, Patterson, H. C., & Page, S. Domain-Referenced Curriculum Evaluation: A Technical Handbook and a Case Study from the MINNEMAST Project. Los Angeles: UCLA, Center for the Study of Evaluation, 1973.

Insgroup, Inc. Excerpts from objectives for instructional programs. Orange, Calif.: Insgroup, Inc., 1972.

Instructional systems development course. Fort Sill, Okla.: U.S. Army Field Artillery School, January 1973.

Instructor's handbook. Fort Benning, Ga.: U.S. Army Infantry School, September 1967.

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

Jackson, R. Developing criterion-referenced tests. ERIC Clearinghouse on Tests, Measurements and Evaluation, 1970.

Kennedy, B. T. The role of criterion-referenced measures within the total evaluation process. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, 1972.

Klein, S.P. and Kosecoff, J. Issues and Procedures in the Development of Criterion-Referenced Tests. Princeton, New Jersey: Educational Testing Service, ERIC TM Report 26, September, 1973.

Knipe, W. H. Diagnostic criterion-referenced testing. Paper presented at the Fall Administrator Meeting, Grand Forks, North Dakota, 1973.

Lindvall, C. M. (Ed.) Defining Educational Objectives. Pittsburgh: University of Pittsburgh Press, 1964.

Livingston, S.A. A classical test-theory approach to criterion-referenced tests. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, 1972. (a)

Livingston, S.A. A reply to Harris' an interpretation of Livingston's reliability coefficient for criterion-referenced tests. Journal of Educational Measurement, 1972, 9, No. 1, 3. (b)

Livingston, S.A. Criterion-referenced applications of classical test theory. Journal of Educational Measurement, 1972, 9(1), 13-26. (c)

Livingston, S.A. The reliability of criterion-referenced measures. Paper presented at the Annual Meeting of the American Educational Research Association, 1971.

Lord, R. M. Estimating norms by item sampling. Educational and Psychological Measurement, 1962, 22, 259-267.

Lyons, J. D. Frameworks for Measurement and Quality Control. HumRRO Professional Paper 16-72, 1972.

Mager, R. F. Preparing instructional objectives. San Francisco: Fearon, 1962

Mager, R. F. Measuring instructional intent. San Francisco: Fearon, 1973.

Marks, E. & Noll, G. A. Procedures and criteria for evaluating reading and listening comprehension tests. Educational and Psychological Measurement, 1967, 27, 339-345.

McFann, H. H. Content Validation of Training. HumRRO Professional Paper 8-73, 1973.

Meredith, K. E. & Sabers, D. L. Using item data for evaluating criterion-referenced measures with an empirical investigation of index consistency. Paper presented at the Annual Meeting of the Rocky Mountain Psychological Association, Albuquerque. 1972.

Miller, R. B. Task description and analysis. In R. M. Gagne (Ed.) Psychologocal Principles in System Development. New York: Holt, Rinehart, and Winston, 1962.

Millman, J. Psychometric characteristics of performance tests of teaching effectiveness. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, February, 1973.

Nitko, A. J. A model for criterion-referenced tests based on use. Paper presented at the Annual Meeting of the American Educational Research Association, New York, 1971.

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

Oakland, T. An evaluation of available models for estimating the reliability and validity of criterion-referenced measures. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, 1972.

Operational policies and procedures. Fort Knox, Ky.: U.S. Army Armor School, November 1973.

Osborn, W. C. An Approach to the Development of Synthetic Performance Tests for Use in Training Evaluation. HumRRO Professional Paper 30-70, 1970.

Osborn, W. C. Developing Performance Tests for Training Evaluation. HumRRO Professional Paper 3-73, 1973.

Osborn, W.C. Process versus product measures in performance testing. Paper presented at the Annual Conference of Military Testing Association, San Antonio, October, 1973.

Osburn, H.G. Item sampling for achievement testing. Educational and Psychological Measurement, 1968, 28, 85-104.

Ozenne, D. G. Toward an evaluative methodology for criterion-referenced measures: Test sensitivity. ED 061263, 1971.

Pieper, W. J. & Swezey, R. W. Learner centered instruction (LCI): Description and evaluation of a systems approach to technical training. Catalog of Selected Documents in Psychology, Spring 1972, 2, 85-86.

Pieper, W. J., Catrow, E. J., Swezey, R. W., & Smith, E. A. Automated apprenticeship training (AAT): A systematized audio-visual approach to self-paced job training. Catalog of Selected Documents in Psychology, Winter 1973, 3, 21.

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

Popham, W. J. Indices of adequacy for criterion-referenced test items. Paper presented at the Annual Meeting of the American Educational Research Association, Minneapolis, 1970.

Popham, W. J. Applications of teaching performance tests to inservice and preservice teacher training. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, 1973.

Prager, B. B., Mann, L., Burger, R. M., & Cross, L. H. Adapting criterion-referenced measurement to individualization of instruction for handicapped children: Some issues and a first attempt. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, 1972.

Preparation of written achievement examinations. Fort Sill, Okla.: U.S. Army Field Artillery School, July 1969.

Rahmlow, H. R., Matthews, J. J., & Jung, S. M. An empirical investigation of item analysis in criterion-referenced tests. Paper presented at the Annual Meeting of the American Educational Research Association, Minneapolis, 1970.

Randall, R. Contrasting norm-referenced and criterion-referenced measures. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, 1972.

Rapp, M.L., Root, J.G., & Summer, G. Some Considerations in the Experimental Design and Evaluation of Educational Innovations. Santa Monica, California: The Rand Corporation, 1970.

Ronan, W. and Prien, E. Toward a criterion theory: A review and analysis of research and opinion. Catalog of Selected Documents in Psychology, 1973, 3, 68.

Roudabush, G.E. Item selection for criterion-referenced tests. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, 1973.

Roudabush, G.E. & Green, D.R. Aspects of a methodology for creating criterion-referenced tests. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, 1972.

Sherman, M., Zieky, M., and Fremer, J. Handbook for conducting task analyses and developing criterion-referenced tests of language skills. Princeton, New Jersey: Educational Testing Service, 1974.

Shoemaker, D.M. Allocation of items and examinees in estimating a norm distribution by item sampling. Journal of Educational Measurement, 1970, 7, 123-128. (a)

Shoemaker, D.M. Item-examinee sampling procedures and associated standard error in estimating test parameters. Journal of Educational Measurement, 1970, 3, 555-562. (b)

Shoemaker, D.M. & Osburn, H.G. Computer-aided item sampling for achievement testing. Educational and Psychological Measurement, 1969, 29, 169-172.

Siegel, S. Nonparametric statistics for the behavioral sciences. H.F. Harlow (ed.) McGraw-Hill Series in Psychology, 1956.

Smith, R.G., Jr. Controlling the Quality of Training. HumRRO Technical Report 65-6, June 1965. AD 618-737.

Smith, R.G., Jr. The Development of Training Objectives. HumRRO Research Bulletin 11, 1964.

Soldiers' Manual Army Testing (SMART). TRADOC Pamphlet No. 600-9, Fort Monroe, Virginia: 1973.

Swezey, R.W. & Pearlstein, R.B. Developing Criterion-Referenced Tests. Reston, Virginia: Applied Science Associates, Technical Report 287-AR18(2)-IR-0974-KWS, 1974.

Systems engineering of training. TRADOC Reg. 350-100-1, Department of the Army, Headquarters, United States Army Training and Doctrine Command, Fort Monroe, Va., 6 July 1973.

Taylor, J. E., Michaels, E. R., & Brennan, M. F. The Concepts of Performance Oriented Instruction Used in Developing the Experimental Volunteer Army Training Program. HumRRO Technical Report TR-73-3, 1973.

Tyler, Ralph W. Constructing Achievement Tests. Ohio State University, Columbus, Ohio, 1934.

Tyler, Ralph W. Basic Principles of Curriculum and Instruction. University of Chicago Press, 1950.

Training handbook. Fort Benning, Ga.: U.S. Army Infantry School.

Tyler, R. W. Some persistent questions on the defining of objectives. Defining educational objectives. C. M. Lindvall (Ed.) University of Pittsburgh Press, 1964.

U.S. Army Infantry School. Training management: An overview, Training Management Digest, No. 1, April 1973. (TC 21-5-1)

U.S. Army Infantry School. Performance-oriented training, Training Management Digest, No. 2, September 1973. (TC 21-5-2, Test Edition)

Vineberg, R. & Taylor, E. N. Performance in four Army jobs by men at different aptitude levels: 4. Relationships between performance criteria. HumRRO Technical Report 72-23, 1972.

Woodson, M.I.C.E. The issue of item and test variance for criterion-referenced tests. Journal of Educational Measurement. 1974, 11(1), 63-64. (a)

Woodson, M.I.C.E. The issue of item and test variance for criterion-referenced tests: A reply. Journal of Educational Measurement. 1974, 11(2), 139-140. (b)

DISTRIBUTION

ARI Distribution List

4 OASD (M&RA)
2 HQDA (DAMI-CSZ)
1 HQDA (DAPE-PBR)
1 HQDA (DAMA-AR)
1 HQDA (DAPE-HRE-PO)
1 HQDA (SGRD-ID)
1 HQDA (DAMI-DOT-C)
1 HQDA (DAPC-PMZ-A)
1 HQDA (DACH-PPZ-A)
1 HQDA (DAPE-HRE)
1 HQDA (DAPE-MPO-C)
1 HQDA (DAPE-DW)
1 HQDA (DAPE-HRL)
1 HQDA (DAPE-CPS)
1 HQDA (DAFD-MFA)
1 HQDA (DARD-ARS-P)
1 HQDA (DAPC-PAS-A)
1 HQDA (DUSA-OR)
1 HQDA (DAMO-RQR)
1 HQDA (DASG)
1 HQDA (DAJO-FI)
1 Chief, Consult Div (DA-OTSG), Adelphi, MD
1 Mil Asst. Hum Res, ODDR&E, OAD (E&LS)
1 HQ USARAL, APO Seattle, ATTN: ARAGP-R
1 HQ First Army, ATTN: AFKA-OI TI
2 HQ Fifth Army, Ft Sam Houston
1 Dir, Army Stf Studies Ofc, ATTN: OAVCSA (DSP)
1 Ofc Chief of Stf, Studies Ofc
1 DCSPER, ATTN: CPS/OCP
1 The Army Lib, Pentagon, ATTN: RSB Chief
1 The Army Lib, Pentagon, ATTN: ANRAL
1 Ofc, Asst Sect of the Army (R&D)
1 Tech Support Ofc, OJCS
1 USASA, Arlington, ATTN: IARD-T
1 USA Rsch Ofc, Durham, ATTN: Life Sciences Dir
2 USARIEM, Natick, ATTN: SGRD-U CA
1 USATTC, Ft Clayton, ATTN: STETC MO A
1 USAIMA, Ft Bragg, ATTN: ATSU-CTD-OM
1 USAIMA, Ft Bragg, ATTN: Marquat Lib
1 US WAC Ctr & Sch, Ft McClellan, ATTN: Lib
1 US WAC Ctr & Sch, Ft McClellan, ATTN: Tng Dir
1 USA Quartermaster Sch, Ft Lee, ATTN: ATSM-TE
1 Intelligence Material Dev Ofc, EWL, Ft Holabird
1 USA SE Signal Sch, Ft Gordon, ATTN: ATSO-EA
1 USA Chaplain Ctr & Sch, Ft Hamilton, ATTN: ATSC-TE-RD
1 USATSCH, Ft Eustis, ATTN: Educ Advisor
1 USA War College Carlisle Barracks, ATTN: Lib
2 WRAIR, Neuropsychiatry Div
1 DLI, SDA, Monterey
1 USA Concept Anal Agcy, Bethesda, ATTN: MOCA-MR
1 USA Concept Anal Agcy, Bethesda, ATTN: MOCA-JF
1 USA Arctic Test Ctr, APO Seattle, ATTN: STEAC-PL-MI
1 USA Arctic Test Ctr, APO Seattle, ATTN: AMSTE-PL-TS
1 USA Armament Cmd, Redstone Arsenal, ATTN: ATSK-TEM
1 USA Armament Cmd, Rock Island, ATTN: AMSAR-TDC
1 FAA-NAFEC, Atlantic City, ATTN: Library
1 FAA-NAFEC, Atlantic City, ATTN: Human Engr Br
1 FAA Aeronautical Ctr, Oklahoma City, ATTN: AAC-44D
2 USA Fld Arty Sch, Ft Sill, ATTN: Library
1 USA Armor Sch, Ft Knox, ATTN: Library
1 USA Armor Sch, Ft Knox, ATTN: ATSB-DI-E
1 USA Armor Sch, Ft Knox, ATTN: ATSB-DT-TP
1 USA Armor Sch, Ft Knox, ATTN: ATSB-CD-AD
2 HQUSACDEC, Ft Ord, ATTN: Library
1 HQUSACDEC, Ft Ord, ATTN: ATEC-EX-E-Hum Factors
2 USAEEC, Ft Benjamin Harrison, ATTN: Library
1 USAPACDC, Ft Benjamin Harrison, ATTN: ATCP-HR
1 USA Comm-Elect Sch, Ft Monmouth, ATTN: ATSN-EA
1 USAEC, Ft Monmouth, ATTN: AMSEL-CT-HDP
1 USAEC, Ft Monmouth, ATTN: AMSEL-PA-P
1 USAEC, Ft Monmouth, ATTN: AMSEL-SI-CB
1 USAEC, Ft Monmouth, ATTN: C, Fact Dev Dr
1 USA Materials Sys Anal Agcy, Aberdeen, ATTN: AMXSY-P
1 Edgewood Arsenal, Aberdeen, ATTN: SAREA-BL-H
1 USA Ord Ctr & Sch, Aberdeen, ATTN: ATSL-TEM-C
2 USA Hum Engr Lab, Aberdeen, ATTN: Library/Dir
1 USA Combat Arms Tng Bd, Ft Benning, ATTN: Ad Supervisor
1 USA Infantry Hum Rsch Unit, Ft Benning, ATTN: Chief
1 USA Infantry Bd, Ft Benning, ATTN: STEBC-TE-T
1 USASMA, Ft Bliss, ATTN: ATSS-LRC
1 USA Air Def Sch, Ft Bliss, ATTN: ATSA CTD ME
1 USA Air Def Sch, Ft Bliss, ATTN: Tech Lib
1 USA Air Def Bd, Ft Bliss, ATTN: FILES
1 USA Air Def Bd, Ft Bliss, ATTN: STEBD-PO
1 USA Cmd & General Stf College, Ft Leavenworth, ATTN: Lib
1 USA Cmd & General Stf College, Ft Leavenworth, ATTN: ATSW-SE-L
1 USA Cmd & General Stf College, Ft Leavenworth, ATTN: Ed Advisor
1 USA Combined Arms Cmbt Dev Act, Ft Leavenworth, ATTN: DepCdr
1 USA Combined Arms Cmbt Dev Act, Ft Leavenworth, ATTN: CCS
1 USA Combined Arms Cmbt Dev Act, Ft Leavenworth, ATTN: ATCASA
1 USA Combined Arms Cmbt Dev Act, Ft Leavenworth, ATTN: ATCACO-E
1 USA Combined Arms Cmbt Dev Act, Ft Leavenworth, ATTN: ATCACC-CI
1 USAECOM, Night Vision Lab, Ft Belvoir, ATTN: AMSEL-NV-SD
3 USA Computer Sys Cmd, Ft Belvoir, ATTN: Tech Library
1 USAMERDC, Ft Belvoir, ATTN: STSFB-DQ
1 USA Eng Sch, Ft Belvoir, ATTN: Library
1 USA Topographic Lab, Ft Belvoir, ATTN: ETL TD-S
1 USA Topographic Lab, Ft Belvoir, ATTN: STINFO Center
1 USA Topographic Lab, Ft Belvoir, ATTN: ETL GSL
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: CTD MS
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: ATS-CTD-MS
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: ATSI-TE
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: ATSI-TEX-GS
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: ATSI-CTS-OR
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: ATSI-CTD-DT
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: ATSI-CTD-CS
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: DAS/SRD
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: ATSI-TEM
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: Library
1 CDR, HQ Ft Huachuca, ATTN: Tech Ref Div
2 CDR, USA Electronic Prvg Grd, ATTN: STEEP MT-S
1 HQ, TCATA, ATTN: Tech Library
1 HQ, TCATA, ATTN: AT CAT-OP-Q, Ft Hood
1 USA Recruiting Cmd, Ft Sheridan, ATTN: USARCPM-P
1 Senior Army Adv., USAFAGOD/TAC, Elgin AF Aux Fld No. 9
1 HQ, USARPAC, DCSPER, APO SF 96558, ATTN: GPPE-SE
1 Stimson Lib, Academy of Health Sciences, Ft Sam Houston
1 Marine Corps Inst., ATTN: Dean-MCI
1 HQ, USMC, Commandant, ATTN: Code MTMT
1 HQ, USMC, Commandant, ATTN: Code MPI-20-28
2 USCG Academy, New London, ATTN: Admission
2 USCG Academy, New London, ATTN: Library
1 USCG Training Ctr, NY, ATTN: CO
1 USCG Training Ctr, NY, ATTN: Educ Svc Ofc
1 USCG, Psychol Res Br, DC, ATTN: GP 1/62
1 HQ Mid-Range Br, MC Det, Quantico, ATTN: P&S Div

1 US Marine Corps Liaison Ofc, AMC, Alexandria, ATTN: AMCGS–F
1 USATRADOC, Ft Monroe, ATTN: ATRO–ED
6 USATRADOC, Ft Monroe, ATTN: ATPR–AD
1 USATRADOC, Ft Monroe, ATTN: ATTS–EA
1 USA Forces Cmd, Ft McPherson, ATTN: Library
2 USA Aviation Test Bd, Ft Rucker, ATTN: STEBG–PO
1 USA Agcy for Aviation Safety, Ft Rucker, ATTN: Library
1 USA Agcy for Aviation Safety, Ft Rucker, ATTN: Educ Advisor
1 USA Aviation Sch, Ft Rucker, ATTN: PO Drawer O
1 HQUSA Aviation Sys Cmd, St Louis, ATTN: AMSAV–ZDR
2 USA Aviation Sys Test Act., Edwards AFB, ATTN: SAVTE–T
1 USA Air Def Sch, Ft Bliss, ATTN: ATSA TEM
1 USA Air Mobility Rsch & Dev Lab, Moffett Fld, ATTN: SAVDL–AS
1 USA Aviation Sch, Res Tng Mgt, Ft Rucker, ATTN: ATST–T–RTM
1 USA Aviation Sch, CO, Ft Rucker, ATTN: ATST–D–A
1 HQ, DARCOM, Alexandria, ATTN: AMXCD–TL
1 HQ, DARCOM, Alexandria, ATTN: CDR
1 US Military Academy, West Point, ATTN: Serials Unit
1 US Military Academy, West Point, ATTN: Ofc of Milt Ldrshp
1 US Military Academy, West Point, ATTN: MAOR
1 USA Standardization Gp, UK, FPO NY, ATTN: MASE–GC
1 Ofc of Naval Rsch, Arlington, ATTN: Code 452
3 Ofc of Naval Rsch, Arlington, ATTN: Code 458
1 Ofc of Naval Rsch, Arlington, ATTN: Code 450
1 Ofc of Naval Rsch, Arlington, ATTN: Code 441
1 Naval Aerospc Med Res Lab, Pensacola, ATTN: Acous Sch Div
1 Naval Aerospc Med Res Lab, Pensacola, ATTN: Code L51
1 Naval Aerospc Med Res Lab, Pensacola, ATTN: Code L5
1 Chief of NavPers, ATTN: Pers-OR
1 NAVAIRSTA, Norfolk, ATTN: Safety Ctr
1 Nav Oceanographic, DC, ATTN: Code 6251, Charts & Tech
1 Center of Naval Anal, ATTN: Doc Ctr
1 NavAirSysCom, ATTN: AIR–5313C
1 Nav BuMed, ATTN: 713
1 NavHelicopterSubSqua 2, FPO SF 96601
1 AFHRL (FT) Williams AFB
1 AFHRL (TT) Lowry AFB
1 AFHRL (AS) WPAFB, OH
2 AFHRL (DOJZ) Brooks AFB
1 AFHRL (DOJN) Lackland AFB
1 HQUSAF (INYSD)
1 HQUSAF (DPXXA)
1 AFVTG (RD) Randolph AFB
3 AMRL (HE) WPAFB, OH
2 AF Inst of Tech, WPAFB, OH, ATTN: ENE/SL
1 ATC (XPTD) Randolph AFB
1 USAF AeroMed Lib, Brooks AFB (SUL–4), ATTN: DOC SEC
1 AFOSR (NL), Arlington
1 AF Log Cmd, McClellan AFB, ATTN: ALC/DPCRB
1 Air Force Academy, CO, ATTN: Dept of Bhl Scn
5 NavPers & Dev Ctr, San Diego
2 Navy Med Neuropsychiatric Rsch Unit, San Diego
1 Nav Electronic Lab, San Diego, ATTN: Res Lab
1 Nav TrngCen, San Diego, ATTN: Code 9000–Lib
1 NavPostGraSch, Monterey, ATTN: Code 55Aa
1 NavPostGraSch, Monterey, ATTN: Code 2124
1 NavTrngEquipCtr, Orlando, ATTN: Tech Lib
1 US Dept of Labor, DC, ATTN: Manpower Admin
1 US Dept of Justice, DC, ATTN: Drug Enforce Admin
1 Nat Bur of Standards, DC, ATTN: Computer Info Section
1 Nat Clearing House for MH- Info, Rockville
1 Denver Federal Ctr, Lakewood, ATTN: BLM
12 Defense Documentation Center
4 Dir Psych, Army Hq, Russell Ofcs, Canberra
1 Scientific Adver, Mil Bd, Army Hq, Russell Ofcs, Canberra
1 Mil and Air Attache, Austrian Embassy
1 Centre de Recherche Des Facteurs, Humaine de la Defense Nationale, Brussels
2 Canadian Joint Staff Washington
1 C/Air Staff, Royal Canadian AF, ATTN: Pers Std Anal Br
3 Chief, Canadian Def Rsch Staff, ATTN: C/CRDS(W)
4 British Def Staff, British Embassy, Washington
1 Def & Civil Inst of Enviro Medicine, Canada
1 AIR CRESS, Kensington, ATTN: Info Sys Br
1 Militaerpsykologisk Tjeneste, Copenhagen
1 Military Attache, French Embassy, ATTN: Doc Sec
1 Medecin Chef, C.E.R.P.A.–Arsenal, Toulon/Naval France
1 Prin Scientific Off, Appl Hum Engr Rsch Div, Ministry of Defense, New Delhi
1 Pers Rsch Ofc Library, AKA, Israel Defense Forces
1 Ministe is van Defensie, DOOP/KL Afd Sociaal Psychologische Zaken, The Hague, Netherlands