DOCUMENT RESUME

ED 178 548 **TM 007 839**

AUTHOR	Swezey, Robert W.; And Others
TITLE	Criterion-Referenced Testing: A Discussion of Theory and Practice in the Army.
INSTITUTION	Army Research Inst. for the Behavioral and Social Sciences, Arlington, Va.
REPORT NO	ARI-RM-75-11
PUB DATE	Dec 75
NOTE	95p.; Appendices marginally legible
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	*Criterion Referenced Tests; Evaluation Needs; Mastery Tests; *Military Training; Needs Assessment; Research Needs; *Test Construction; *Testing; Testing Problems; Use Studies
IDENTIFIERS	*Army

ABSTRACT

As the basis for developing a criterion referenced test (CRT) construction manual for the Army and for identifying potential research areas, a study was conducted which included a review of the technical and theoretical literature on criterion referenced testing and a survey of CRT applications at selected Army installations. It was found that the use of CRT's was limited, although some serious attempts were being made to develop and administer them. Progress was noted in such areas as equipment related skills, but little evidence of CRT development was found in "soft skill" areas or in team performance situations. There was general consensus that clearly-written CRT construction guides were needed. Difficulties were observed in CRT development and use: lack of task analysis data and well-defined objectives; inattention to prioritizing tasks; disregard for practical constraints; insufficient number of items in the item bank for alternate test forms and lack of item analysis techniques; omission of test reliability and validity studies; and lack of standardized testing conditions. (A lengthy bibliography and appendices, including the interview form, summary of types of individuals interviewed, and quantitative data gathered at each installation, are provided). (MH)

Research Memorandum 75-11

CRITERION-REFERENCED TESTING: A DISCUSSION OF THEORY AND PRACTICE IN THE ARMY

Robert W. Swezey, Richard B. Pearlstein and William H. Ton
Applied Science Associates, Inc.

UNIT TRAINING AND EVALUATION SYSTEMS TECHNICAL AREA

U. S. Army

Research Institute for the Behavioral and Social Sciences

December 1975

"APPROVED FOR PUBLIC RELEASE, DISTRIBUTION UNLIMITED."

Army Project Number
2Q164715A757

Unit Training Standards
and Evaluation

Research Memorandum 75-11

CRITERION-REFERENCED TESTING: A DISCUSSION OF THEORY AND OF PRACTICE IN THE ARMY

Robert W. Swezey, Richard B. Pearlstein and William H. Ton
Applied Science Associates, Inc.

Angelo Mirabella, Work Unit Leader

Submitted by:
Frank J. Harris, Chief
UNIT TRAINING AND EVALUATION SYSTEMS TECHNICAL AREA

December 1975

Approved by:

Joseph Zeidner, Director
Organizations and Systems
Research Laboratory

J. E. Uhlaner, Technical Director
U.S. Army Research Institute for
the Behavioral and Social Sciences

CRITERION-REFERENCED TESTING: A DISCUSSION OF THEORY AND PRACTICE IN THE ARMY

CONTENTS

TABLES (continued) Page

CRITERION-REFERENCED TESTING: A DISCUSSION OF THEORY AND PRACTICE IN THE ARMY

INTRODUCTION

This report is an interim document dealing with development of a Criterion-Referenced Test (CRT) Construction Manual. The major objectives of the study were the development of an easy-to-use, "how-to-do-it" manual to assist Army test developers in the construction of CRTs, and the identification of needed research to help achieve a more consistent, unified criterion-referenced test model.

In order to accomplish these objectives, the study: surveyed the literature on criterion-referenced testing in order to provide an information base for development of the CRT Construction Manual; visited selected Army posts to review the present status of criterion-referenced test construction and application in the Army; prepared a draft CRT construction manual; conducted a trial application of the draft manual; and revised the CRT construction manual. The manual for developing criterion-referenced tests has been published as an ARI Special Publication: Guidebook for Developing Criterion-Referenced Tests.

Part 1 of this report reviews the technical and theoretical literature in criterion-referenced testing. This review is a serious discussion of the state-of-the-art in criterion-referenced testing, designed for the academically-oriented reader. The review discusses questions of CRT reliability and validity in both practical and theoretical areas, different methods of CRT construction, simulation fidelity (e.g., the extent to which CRTs can and should mirror real-world performance conditions), the use of CRTs in mastery learning contexts and to test development and item sampling, diagnostic uses of CRTs, the establishment of passing scores, and uses of CRTs in public education and military contexts.

Part 2 describes a survey of Army CRT applications at a number of Army installations. Results of the survey are indicated through an analysis of quantitative data collected during interviews and through a discussion of qualitative comments received, problems observed, and areas where changes may prove beneficial to the Army.

Appendices A, B, and C provide, respectively, the Interview Protocol used during the Army CRT survey; a summary of types of individuals interviewed at each Army installation surveyed; and quantitative data gathered at each Army post.

PART 1--REVIEW OF TECHNICAL AND THEORETICAL LITERATURE

Criterion-referenced testing (CRT) has been widely discussed since the term was popularized by Robert Glaser in 1963. In CRT, questions involving comparisons among individuals are largely irrelevant. CRT information is usually used to evaluate the student's mastery of instructional objectives, or to approximately locate him for future instruction (Glaser and Nitko, 1971). A CRT has been defined variously in the literature, in fact definitions vary so widely that a given test may be classified as either a CRT or a norm-referenced test (NRT) according to the particular definition used. Glaser and Nitko (1971) propose a flexible definition:

> "A CRT is one that is deliberately constructed so as to yield measurements that are directly interpretable in terms of specified performance standards.... The performance standards are usually specified by defining some domain of tasks that the student should perform. Representative samples of tasks from this domain are organized into a test. Measurements are taken and are used to make a statement about the performance of each individual relative to that domain."

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. Lyons (1972) argues for the use of criterion-referenced measurement 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."

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

The contemporary interest is mastery learning has led to a growing interest in CRT. CRTs can be used to serve 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 as to "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 of the effectiveness of instruction because they are not derived from the particular task objectives. CRT is, however, useful for the evaluation of instruction because of the specificity of the results to the task objectives (Lord, 1962; Cronbach, 1963; Shoemaker, 1970a, 1970b; Hambleton, Rovinelli, and Gorth, 1971).

Popham (1973) points 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) poses a series of arguments against the use of CRT in education. Ebel points out with some justification that CRT measures do not tell us all we need to know about educational achievement, pointing out that CRT measures are not efficient at discovering relative strengths and deficiencies. This is true and is an excellent case for combining CRT with NRT in cases where both relative and absolute information must be gathered. Ebel also raises an objection shared by many practicing educators to the whole "systems" approach to educational development. That is, objectives specific enough to support the generation of CRT are more likely to suppress than to stimulate "good teaching". Ebel leaves us, however, without a metric capable of defining "good teaching" and the untenable assumption that "good teaching" is the rule. Finally, Ebel confuses the concept of mastery of material with the practice of using percentile grades as pass-fail measures. Ebel does not address the notion that CRT as currently constructed are the result of the application of a carefully thought out analysis and development system.

RELIABILITY AND VALIDITY

As Glaser and Nitko (1971) point out, the appropriate technique for an empirical estimation of CRT reliability is not clear. 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. CRTs must be internally consistent and stable, yet estimates of indexes that are dependent on score variability may not reflect this. This section will critically examine a number of studies which have addressed the question of reliability. The question of validity of CRTs 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.

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 on the post-test. Also a DI was obtained in the more conventional manner. After post-test, the distribution of scores was divided into the upper third and the lower third, then the percentage of students in the lower third was subtracted from the percentage of students in the upper third. The Spearman Rho's obtained between the two DI's 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 Gorth (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, however, this method uses statistical procedures dependent on score variability which are questionable for CRT (Popham and Husek, 1969; Randall, 1972) particularly if it is to be employed for item selection (Oakland, 1972).

Livingston (1972a) acknowledges Popham and Husek's comment that "the typical indexes of internal consistency are not appropriate for criterion-referenced tests". Nevertheless, Livingston feels that the classical theory of true and error scores can be used in determining CRT reliability. Livingston 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 deviation 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 negative variety can be minimized by accepting as true masters the group that comfortably exceeds the required criterion level. 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 the group of students. This provides an effective limitation for the computation of inter-item correlations as it is often difficult to ensure equal difficulty levels, which must fluctuate with the group being tested.

Regarding Livingston's (1972a) proposal that the psychometric theory of true and error scores could be adapted to CRT, Oakland (1972) commented that the procedures seemed viable but that the conditions under which they could be used were overly restrictive.

Harris (1972) objects 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)

indicates that Harris overlooked the point that reliability is not a property of a single score but of a group of scores. Livingston also points out 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 take issue with Livingston's concept of CRT reliability estimation as variability around the criterion score, pointing out that CRT is concerned primarily with the accuracy of the pass-fail decision and is relatively unconcerned with a person's attainment above or below the criterion level.

Roudabush and Green (1972) present an analysis of false positive and false negative to derive reliable estimates. These authors presented several methods for arriving at reliability estimates for CRT. The first involves ordering items into a hierarchical order of increasing difficulty. Roudabush and Green propose that error of measurement would be demonstrated if a student failed an easier item while passing a series of more difficult items. Oakland (1972) points out that it is exceedingly 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 propose a second technique utilizing point-biserial correlation between parallel tests. Their results with this method were far from encouraging. In addition, there is great difficulty inherent in the development of parallel tests. The third method involves the use of regression equations to predict item criterion scores but has not yet been fully explored.

In a divergent work, Hambleton and Novick (1971) propose 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 and properly so, as the severly restricted variance encountered with CRT will cause correlationally-based estimates of reliability and validity to be artificially low. Hambleton and Novick view a decision theoretic metric such as a "loss function" as being more appropriate for use on CRTs. This metric must serve to describe if an individual's true score is above or below a cutting score. The concept differs markedly from Livingston's (1972a) notion in which the criterion is regarded as the true score.

The importance of correct decision-making in CRT applications is also recognized by Edmonston, Randall, and Oakland (1972) who present a CRT reliability model aimed at supporting decisions made during formative evaluation and maximizing the probability of learning an established set of objectives. Criterion-referenced items are usually binary 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. recommend utilizing the cell proportions to provide information about the relationships between the variables represented by the table. They find 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). They also recommend a supplemental measure λ_r (Lambda) a variance-free coefficient. Goodman and Kruskal (1954) define λ_r

$$\lambda_r = \frac{\Sigma p_{aa} - \frac{1}{2}(PM' + P'M)}{1 - \frac{1}{2}(PM' + P'M)}$$

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 goind from a no-information situation to the other-method-known situation. Edmonston et al. feel the reliability estimate most useful to CRT is the extent to which they fluctuate temporally. They fell that, minimally, CRT items should provide stable estimates of knowledge of curriculum content; $\Sigma_a p_{aa}$ and r can be used to provide estimates of this stability. They recommend that $\Sigma_a p_{aa}$ 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 90%) 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. If knowledge of the re-test score provides additional information, the item is retained. However, there is no current basis for determining the acceptable minimal reduction in classification error.

In the same vein as Edmonston et al., Roudabush (1973) views reliability as referring to the appropriateness of the decisions made that affect the treatment of the examinee. Roudabush emphasizes "Minimizing risk or cost to examinee." The decision iw whether to discontinue instruction or remediate or wash-out.

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

McFann (1973) views the content validation of training as having two major dimensions. The first dimension is the role of the human within the general operating system. Generally, this is defined by means of task analysis. The second dimension involves the skills and knowledge the trainee brings with him to the course; the 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 the training must also be tempered by management orientation to cost and effectiveness. Finally, McFann feels that

decisions made on the units or procedures by which output is to be evaluated has an influence on validation of training content. McFann views the validation of training content as a dynamic, interactive process whereby training content is initially determined and then, on the basis of feedback of student performance on the job, instructional content as well as instruction method is modified to improve overall system effectiveness.

Edmonston, Randall, and Oakland (1972) hold content validation as central to CRT development. "CRT items are sampled theoretically from a large item domain and must be representations of a specified behavioral objective.

Hambleton and Novick (1971) propose a validity theory in which a new test Y would serve as criterion. The qualifying score of the second test need not correspond with the qualifying score of the predictor CRT. The Test Y these authors suggest might be derived from performance on the next unit of instruction, or it may be a job-related performance criterion. Although this appears to be a good idea, it seems that different conclusions would be reached if test Y were a job-related criterion instead of performance on the next unit of instruction. The fact that the conclusion might be different could, however, yield an approximation 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 in the case where the tasks on the two tests are similar.

Edmonston et al. (1972) advocate a method of CRT validation which they term the criterion-oriented approach, which 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 and training observers as well as performance on suitable NRT measures). Edmonston et al. view these as measures of concurrent validity, although these multiple indicators could, if properly chose, provide an estimate of construct validity. In addressing the problems of predictive validation, Edmonston et al. concur with Kennedy (1971) in 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. This notion has been partially verified by Edmonston and his co-workers. 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. advocate an approach to construct validity initially put forth by Nunnally (1967). 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 not too far from Cronback and Meehl's (1955) enunciation of the need for a "nomological network" with which to validate a construct. Edmonston et al. indicate that the "specification of a hierarchy of learning sets among items would seen 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 the establishment of a network of relationships sufficient to completely define a construct.

In Roudabush's (1973) view of validity, CRT items are designed to sample as purely as possible the specified domain of behavior, then tried out to determine primarily if the items are sensitive to instruction. A 2x2 contingency table containing post-test and pre-test outcomes is the basis for analysis:

Pre-test \ Post-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) assume f_3 due to guessing and derive a sensitivity index(s) that is simply the proportion of cases that missed the item on the pre-test and passed it on the post-test with a correction for guessing.

$$s = \frac{\hat{f}_2}{\hat{f}_1 + \hat{f}_2} \quad \text{where}$$

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

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

Roudabush (1973), however, found that to derive a "reasonably reliable" value for the index there should be 50 cases who missed the item at pre-test (f_1), while if f_4 cell is high the index will have little value (neither will the item). This index ranges from 1.00 to 0.00 but may go below 0.0 if miskeyed. 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 contemplated.

These various treatments of CRT validity all exhibit difficulties that often might prove insurmountable to a test constructor dealing with "real world" problems. Content validity, however, is extremely important in CRT 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 to demonstrate this form of validity. Predictive validity appears practicable in many situations.

CONSTRUCTION METHODOLOGY

NRTs are primarily designed to measure individual differences. The meaning which can be attached to any particular score depends on a comparison of that score to a relevant norm distribution. A norm-referenced test is constructed specifically to maximize the variability of test scores since such a test is more likely to produce fewer errors in ordering the individuals on the measured ability. Since NRTs are often used for selection and classification purposes, it follows that minimizing the number of order errors is extremely important.

NRTs are constructed using traditional item analysis procedures. It is partly because of this that the test scores cannot be interpreted relative to some well-defined content domain since items are normally selected to produce tests with desired statistical properties (e.g., difficulty levels around .5), rather than to be representative of a content domain. Likewise, a wide range of item difficulty does not occur because of resulting variance restriction. 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 with the responses of others. There is no interest in absolute measurement of individual skills as in CRTs, only relative comparison.

Although conceptually allied to the construction of NRTs, item analysis is an important tool in assembling a test from an item pool and therefore has application to the construction of certain CRTs. Although content validity is an important characteristic for an item in a CRT, there are other important considerations having to do with the sensitivity and discriminating power of an item. These features are important when evaluating instruction and in ensuring the correct decision regarding an individual's progress through instruction.

In CRT development, the item difficulty index is useful for selecting "good" items. However, item difficulty is used differently than in NRT. If the content domain is carefully specified, test items written to measure accomplishment of the objectives should also be carefully specified and

closely associated with the objectives. Therefore, all of the items associated with the same objectives should be answered correctly by about the same proportion of examinees in a group. Items which differ greatly should be carefully examined to determine if they coincide with the intent of the objectives.

Similarly, item discrimination indexes can be useful for CRT development. Negative discrimination indexes warn that CRT items need modification, or that the instructional process is at fault. A negative index would be indicative of a high proportion of "false negatives"; conversely a positive discrimination index is useful for diagnosing shortcomings in the instructional program.

An attempt to use item analysis techniques to develop test evaluation indexes was undertaken by Ivens (1970). Ivens defines reliability indexes based on the concept of within S equivalance of scores. Item reliability is defined as the proportion of subjects whose item scores are the same on the post-test and either a re-test or parallel form. Score reliability is then defined as the average item reliability. Unfortunately the need for re-test or for two forms (parallel) would seem to reduce the usefulness of this scheme except in very special situations.

Rahmlow, Matthews and Jung (1970) suggest that the function of a discrimination index in a CRT is primarily that of indicating the homogeneity of the item with respect to the specific instructional objective measured. These authors focus attention on a shift in item difficulty from pre-instruction to post-instruction.

Helmstadter (1972) compared 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.

Helmstadter also sought to compare the traditional item discrimination index applied to pre- and post-instruction with difficulty indexes derived in the same fashion. His findings confirmed that caution should be observed in the use of traditional item analysis procedures in CRT. In a similar finding, Roudabush (1973) showed that use of traditional item statistics would have resulted in some objectives being over-represented while others would be represented by no items.

Ozenne (1971) has developed an elaborate model of subject response which he uses to derive an index of sensitivity. In this formulation the sensitivity of a group of comparable measures given to a sample of S's before and after instruction is the variance due to the instructional effect divided by the sum of the variance due to the instructional effect and error variance. The index was, however, developed for a severely restricted sample 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 for specific cases of CRTs but evidence as to their generalizability is lacking. If item analytic procedures are to be used in evaluating CRTs, then it must be known what sort of score is produced by that item. The usual score is a pass-fail dichotomy. A CRT item can result in two types of incorrect decisions. Roudabush and Green (1972) refer 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 the CRT to make the "right" decision, i.e., avoiding false negatives and false positives. The adequacy of a CRT in these authors' view is determined by its ability to discriminate consistently and appropriately over a large number of items.

Carver (1970) proposed two procedures to assess reliability of a CRT item. For a single form he suggests comparing the percentage meeting criterion level in one group to the same percentage in another "similar" group; for homogeneous sets he recommends using one group and comparing the percentages identified as meeting criterion on all items. Meredith and Sabers (1972) point out, however, that it must be determined how two CRT items, whether identical or parallel, identify the same individual with regard to his attainment of criterion level. With regard to item analysis procedures, if a CRT item is administered before and after instruction, and it does not discriminate, there are alternatives to labeling it unreliable. A non-discriminating item may simply be an invalid measure of the objectives or it may indicate that the instruction itself is inadequate or unnecessary.

Meredith and Sabers suggest the use of a matrix consisting of the pass-fail decisions of two CRTs. By defining the two CRT items as being the same measures we can examine test/re-test reliability, but without time intervening between the measures, the reliability is of the concurrent or internal consistency variety. In addition, undefined problems exist with acceptably defining two CRTs as the same. Various other indexes are possible but a great weight is placed upon carefully defining relationships between measures a priori. Considerable confusion is evidence in the use of "same" and parallel forms without formal definitions. Similarly it is stated that if one CRT item is a "criterion measure", then the validity of the other CRT can be found. By definition, both are criterion measures and if the "criterion measure" is external to the instructional domain, then it is not a CRT item in the same sense. Various coefficients are given but the difficulty in definition mentioned above limits their usefulness.

FIDELITY

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

1. Solicit opinions. This category, the lowest level, man in fact often miss the payoff questions (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 produce 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 the six levels proposed by Frederiksen possesses both advantages and disadvantages. An optimal solution would be to assess individual performance at the highest possible level of fidelity. Unfortunately, deriving performance data may involve a subjective (rating) technique for a specific situation, 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, Root, and Sumner, 1970). An actual increase in overall criterion adequacy may result from a gain in objectivity which may compensate for a corresponding loss in fidelity.

The question of fidelity addresses the issue of how much should the test resemble the actual performance. Fidelity is not usually at issue in NRT and has its primary application in criterion-referenced performance tests. There are trades to be made between fidelity and cost. A more salient issue, however, is how to empirically modify face fidelity to satisfy needs of the testing situation while retaining the essential stimuli and demand characteristics of the real performance situation.

Osborne (1970) addresses problems in finding efficient alternatives to work s mple tests. Osborn was concerned with developing a methodology that would allow derivation of cheaper procedures that would preserve content validity. There are many realistic situations where job sample tests are not feasible, and job-knowledge tests are not relevant. Obviously the existence of intermediate measures would be a great boon to evaluating performance in this situation. However, methods for developing intermediate or "synthetic" measures are lacking. Osborn gives a brief outline of a method for developing these synthetic measures. Osborn presents 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 test method most cost-effective for each behavior. The tradeoff that must be made between test relevance, related diagnostic performance data, and ease of administration and cost is obvious, and must be resolved by the judgement of the test constructor. Osborn's notions are intriguing but much more development is needed before a workable method for deriving synthetic performance tests is available.

Vineberg and Taylor (1972) address a topic allied to the fidelity issue, that is: to what extent can job knowledge tests 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 individual performance is often considered to be poor practice by testing "experts". 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 with an opportunity to examine the relationship between job sample test scores and job knowledge test scores in four U.S. Army jobs that varied greatly in job type and task 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 that information that is 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 skill requirements are determined by careful job analysis to be minimal.

In a similar work, Engel and Rehder (1970) compared peer ratings, a job knowledge test, and a work-sample test. These workers found that while the knowledge test was acceptably reliable, it lacked validity, and reading ability tended to enter into performance. Peer ratings were judged to have unacceptable validity. Ratings were also essentially uncorrelated with the written test. The troubleshooting items on the written test exhibited a moderate but useful level of validity, while the corrective-action items had little validity. Finally, Engel and Rehder note that the work-sample is the most costly method and is difficult to administer, while the peer ratings and written tests were the least costly and were easy to administer.

Osborn (1973) discusses an important topic related to both the validity and fidelity of a CRT. Osborn points out that task outcomes and products are used to assess student performance while measures of how the tasks are done (processes) 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.

Relatively few tasks are of the first type. Osborn offers gymnastic exercises or springboard diving as examples. More tasks are of the second type, i.e., fixed procedure tasks. In these tasks, if the process is correctly executed the product follows. A great many tasks are of the third type where the process appears to have been correctly carried out but the product was not attained. Osborn offers two reasons why this can happen: either, 1) we were unable to specify fully the necessary and sufficient steps in task performance, or 2) because we do not or cannot accurately measure them. An example of aim-firing a rifle is given as an illustration that there is no guarantee of acceptable markmanship even if all procedures are followed. In this case, process measurement would not adequately 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. Osborn, however, discusses a number of type 3 tasks where product measurement is impractical because of cost, danger, or practicality. In these cases process measures would come to be substituted with resulting injury to the validity of the measure. Osborn poses a salient question that the test developer must answer: 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 that that to be expected by errors of measurement, the test developer should pause and reconsider ways in which times and resources could be compromised in achieving at least an approximation to product measurement. Osborn concludes 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 (1978) has also addressed process versus product measurement, and assist versus non-interference methods of scoring in CRT development. Swezey has recommended process measurement in addition to, or instead of, product measurement when: Diagnostic information is desired, when additional scores are needed on a particular task, and when there is no product at the end of the process.

An issue which must be faced when constructing a complex CRT is the bandwidth fidelity problem (Cronback and Gleser, 1965), i.e., the question of whether to obtain precise information about a small number of competencies or less precise about a larger number. Hambleton and Novick (1971) conclude that the problem of how to fix the length of each sub-scale to maximize the 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 NRT is well established and highly specified, the construction of CRT has been much more of an art. There have been, however, several attempts to formalize the construction of CRT. Ebel (1962) describes the development of a criterion-referenced test of 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. To the extent that scores are reproducible on tests developed independently under the same procedures, the scores may be said to have inherent meaning. Flanagan (1962) indicates that a variant of Ebel's procedure was used in project TALENT. The tests used in the areas of spelling, vocabulary, and reading were not based on specific objectives. They were, however, 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 conditions as prerequisites for allowing inferences to be made 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 by random sampling from the content universe.

It is rarely feasible to satisfy the first conditions in any complete fashion for complex behavior domains. However, the problem of testing all items can be overcome at least in a highly specified content area by the use of an item form (Hively, Patterson, and Page, 1968, 1970; Osborn, 1968). The item form generally has the following characteristics (Osborn, 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.

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. However, generalizing these results to other domains has led to the finding that the difficulty of objectively defining a test construction process is 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.

It appears that at the current state-of-the-art, it is difficult to develop the objective procedures necessary for criterion-referenced measurement of complex behavior without doing violence to measurement objectives. What is needed for complex content domains are item generating rules that permit generalizations of practical significance to be made.

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 feels 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 (1975) suggests fourteen steps in the course of developing a test for training evaluation. The first three steps have to do with assembling information concerning the skills and knowledge segments, the relative importance of each objective, and the completeness of each objective. In step 4 the developer should obtain classification concerning measuring of confusing elements. Osborn points out that performance standards are generally a source of trouble. Steps 5-8 concern themselves with developing the test items and answering questions of the feasibility of simulation as well as questions of controlled administration. In step 9, a final aspect of measurement reliability is considered. Here procedures for translating observed performance into a pass-fail score must be developed. Unfortunately, Osborn does not tell us how to develop pass-fail criteria that will generalize to trainees' performance in the field. In step 10 a supplementary scoring procedure is developed for diagnosing reasons for trainee failure. Osborn does not say if this is to be a criterion- or norm-referenced interpretation. In step 11 the developer formats the final item with its instruction, scoring procedures, etc. In step 12 a decision is made as to whether time permits testing on all objectives or if a sample should be used.

Step 13 covers sampling procedures based on the criticality of the behavior. In step 14 guidance for test administration is prepared. Osborn has provided the developer of CRTs with a broad outline of the steps to be taken in item development. Unfortunately, he does not provide much detail on how various decisions are to be made, i.e., what are passing scores, how to simulate, etc. It is the quality of these decisions that determines the usefulness of the final instrument but the decision-making process apparently remains an art.

MASTERY LEARNING

Besel (1973a,b) contends that norm-group performance is useful and legitimate information for the construction and application of CRT. Besel defines a CRT as a set of items sampled from a domain which has been judged to be an adequate representation of an instructional objective. The domain should be fully described so as to allow two test developers to independently generate equivalent items which measure the same content and are equally reliable. A degree of arbitrariness creeps in when a mastery level is specified for a given objective or set of objectives. Besel recommends the "Mastery Learning Test Model" to provide an appropriate algorithm to support mastery/non-mastery decisions. Two statistics are computed: The probability that a student has indeed achieved the objective and the proportion of a group which has achieved the objective. The model assumes that each student can be treated as either having achieved the objective or not having achieved the objective with partial achievement possible. The Mastery Learning Test Model and its underlying true score theory is related to a notion enunciated by Emrick (1971). Emrick assumed that measurement error was attributable to two sources: α, the problem that a non-master will correctly answer an item ("false positive") and β, the probability that a master will give an incorrect answer to an item ("false negative"). These constructs resemble the Type I and Type II errors encountered in discussions of statistical inference. Emrick's model assumes that all item difficulties and inter-item correlations are equal, a difficult assmption in view of the assumed variability of the former as a result of instruction and the difficulties in computing the latter. Besel (1973 a, b) had developed algorithms for estimating α and β. Three data sources are used:

1. Item difficulties
2. Inter-item co-variance
3. Score histograms

In a tryout, Besel reports "that the usage of an independent estimate of the proportion of students reaching mastery resulted in improved stability of Mastery Learning parameters." This improved stability of A and B should promote increased confidence in mastery/non-mastery decision. 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 A. 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 the competing errors of misclassification. These decisions may have to be made individually for each instructional situation.

ESTABLISHING AND CLASSIFYING INSTRUCTIONAL OBJECTIVES

The development of student performance objectives for instructional programs has become a widespread and well-understood process throughout the educational community. For quality control of the conventional process crucial information derives directly from instructional objectives; they provide not only the specifications for instruction, but also the basis for evaluating instruction (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 the goals of education in terms of what would be meaningful and useful to the classroom teacher. Tyler's work has had considerable impact in the trend toward describing objectives in terms of instructional outcomes.

The second development has come from the need to specify man-machine interaction in modern defense equipment. Miller (1962) was responsible for pioneering efforts in developing methods for describing and analyzing job tasks. Chenzoff (1964) reviewed the then exact 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, i.e., knowledge, comprehension, etc.

2. Task analysis based upon behavioral analysis, i.e., chains, concepts, etc.

3. Task analysis based on information processing needs for performance, i.e., indicators, uses, etc.

4. Task analysis based on a decision paradigm which emphasizes the judgement and decision-making rationale of the task.

5. Task analysis based upon subject matter structure of a task.

6. Task analysis based upon vocational schematics which involve analysis of jobs, duties, tasks and task elements.

The point of Davies' breakdown is that there is no <u>one</u> task analysis procedure. The general approach is to "gin up" a new task analysis scheme or modify an existing scheme to suit the needs of the job at hand.

The third development was the concept of programmed instruction which required the writers of programs to acquire specific information in instructional objectives.

It is apparent that these initial phases of development have largely merged, and the use of instructional objectives has become accepted educational practice. A critical event in this fusion was the publication of Mager's (1962) little book Preparing Instructional Objectives. In this work, Mager set forth the requirements for the form of a useful objective but he did not deal with the procedures by which one could obtain the information to support preparation of the objectives. A series of additional works including one on measuring instructional intent (Mager, 1973) have dealt more thoroughly with such issues.

Information as to the actual behaviors exhibited by an acceptable performer is preferred as the basis for the construction of an instructional objective. However, data can come from a variety of sources, such as:

1. Supervisor interview
2. Job incumbent interview
3. Observation of performer
4. Inferences based on system operation
5. Analysis of "real world" use of instruction
6. Instructor interview

The methods used to derive this data are legion and have become very clever and sophisticated. Flanagan's (1949) "critical incident technique" and the various modifications and off-shoots it has inspired is a good example of an effort aimed at identifying essential performance while eliminating information not directly related to the successful accomplishment of a job-related task.

The choice of method for deriving job behavior instruction must be based on the type of performance and various realistic factors such as the 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 the objectives so derived and provide a useful critique of the data collection method. An exhaustive review of the various techniques for deriving instructional objectives is impossible here. The reader is directed to Lindvall (1964) and Smith (1962) for a comprehensive treatment of this question.

Ammerman and Melching (1966) have developed a system for the analysis and classification of terminal performance objectives. Ammerman and Melching examined a great number of objectives generated by different agencies 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

Further, Ammerman and Melching have identified a number of levels under each factor. 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 the identification of levels for each factor, it is possible to classify or code any terminal objective by a five digit number. This scheme has high value for management control and 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 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 an extremely useful method, particularly where a panel of judges is used to review each objective. A coefficient of congruence can be computed between the judge's placement of the objective on the five dimensions to yield a relative index of agreement. Used in this fashion, the Ammerman and Melching method should prove to be very useful in development of instructional systems.

DEVELOPING TEST MATERIALS AND ITEM SAMPLING

Hively and his associates (1968, 1973) provide a useful scheme for writing items which are congruent with a criterion. Hively's effort has been in the area of domain-referenced achievement testing. In Hively's 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) points out that this approach has met with success where the content area has well-defined limits. In areas such as mathematics, independent judges tend to agree on whether a given item is congruent with the highly specific behavior domain-referenced by the item form. As less well-defined fields are approached, however, it becomes very difficult to prepare item forms so that they yield test items which can be subsequently judged congruent with a given instructional objective. Easy interjudge agreement tends to fade

and the items become progressively more cumbersome. Popham (1970) remarks: "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 presents 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 according to a plan. Analysis of variance techniques (particularly intra-class correlation) are used to obtain estimates of components of variance due to sampling error, testing conditions, and other sources which may affect the reliability of the score. It should be pointed out that analysis of variance, when used in this fashion, is essentially a non-parametric technique particularly suitable for use with CRTs. Generalizability theory has been extended (Osburn, 1968) by including the concepts of task analysis which allows sorting subject matter into well-defined behavioral classes. Osburn (1968) has termed this convergence "Universe-defined achievement testing". Hively et al, (1968, 1973) has used these techniques in an exploration of the mathematics curriculum. Mathematics represents a subject domain particularly suited to this approach and Hively reported success as evidenced by him in the high intra-class correlations between sets of items sampled from a universe of items. If applicable to less well-defined content domains, this technique promises to have diagnostic utility and also particular relevance to examining the form of relationships between knowledges and skills. As yet, this extension into other subjects has not been undertaken

QUALITY ASSURANCE

In the view of Hanson and Berger (1971) quality assurance is viewed as a means for maintaining desired performance levels during the operational use of a large scale instructional program. These workers identify six major components in a Quality Assurance program:

1. Specification of indicator variables. These are variables which measure the important attributes of aspects of a program and must be individually defined for each instructional system.

Examples given are:

a. Pacing--measure of instructional time

b. Performance--interim measures of learning, i.e., unit tests, module tests, etc.

c. Logistics--indicator reports of failure to deliver materials, etc.

2. **Definition of decision rules. The emphasis here should be on** indicators which signal a major program failure. Critical levels may be determined on the basis of evidence from developmental work or on the basis of an analysis of program needs.

3. Sampling procedures. These questions must be answered on the basis of an analysis of the severity of effects if sufficient information is available. 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 here 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 without special effort.

5. Analysis and summarization of data. Some data may be analyzed as it comes 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 must describe the actions to be taken in the event of major program failure. Alternatives should be generated and scaled to the severity of the failures. Information as to actions taken to correct program failures should always be fed back into the program development cycle. This feedback will be an important source of information to guide program revision.

Hanson and Berger offer an illustrative example of how this process might be implemented. They conclude by noting that quality assurance, as applied to criterion-referenced programs, would act to ensure that the specified performance levels will be maintained through the life of a program. These notions provide the basis of an important concept in the implementation of an instructional program utilizing criterion-referenced measurement. If this sort of internal quality assurance program is built into the instruction, then the probability of an instructional program becoming "derailed" while up and functioning is certainly minimized.

DESIGNING FOR EVALUATION AND DIAGNOSIS

Baker (1972) feels that the critical factor in instruction is not how the test results are portrayed (NRT or CRT) but how 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 range. These tests are results of the norm-referenced type. Criterion-referenced tests, Baker feels, are probably better termed domain-referenced tests (see dicsussion of Hively et al., 1968, 1973). A domain specifies both the performance the learner is to demonstrate as well as the content domain to which the performance is to generalize. Another subset of CRT is what Baker refers to as the objective-referenced test. The objective-referenced test starts with an objective based on observable behavior from which it is possible to produce items which are homogeneous yet relate to the objective. Baker feels the notion of domain-referenced tests is more useful.

Each type of test will provide different information to guide improvement on instructional systems. Construct-referenced tests will provide information regarding a full range of content and behavior relevant to a particular construct. The objective-referenced test will provide items which exhibit similar response requirements relating to a vaguely defined content area. The domain-referenced test will 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.

Baker (1972) then proposes a minimum set of data needed to implement as instructional improvement cycle.

1. Data on applicable student abilities
2. Ability to identify deficiencies in student achievement
3. Ability to identify possible explanation for deficiencies
4. Ability to identify alternative remedial sequences
5. Ability to implement sequence

All three types of tests provide data useful for set 1. Construct-referenced tests are probably the most readily available, but are not administered on a cycle compatible with diagnosis and are reported in a nomothetic manner. A well-designed objective-referenced test may be scheduled in a more useful fashion. A domain-referenced test provides enabling information to allow instructors to identify what the students were able to deal with. Identification of performance deficiencies (set 2) is theoretically possible with all three sets of data. However, since cut-offs are usually arbitrary, none of the three tests will give adequate information.

As for sets 3, 4, and 5, there is little in the way of information yielded by any of the three tests which would aid in these decisions. In addition, training research is not yet well-advanced in these areas, nor does the information always reach the user level. In addition, incentives are lacking since most accountability programs are used to punish deficiency rather than to promote efficienty. Of the three test types, the domain-referenced tests give program developers the most assistance, for they are provided with clear information about what kind of practice items are in the area of content and performance measured by the test. Also students may practice on a particular content domain without contacting the test items themselves. However, Baker points out domain-referenced items are hard to prepare, mainly because not all content areas are analyzed in a fashion to allow specification of the behaviors in the domain, as has been noted elsewhere.

ESTABLISHING PASSING SCORES

Prager, Mann, Burger, and Cross (1972) discuss the cut-off point issue and point out that there are two general routes to travel. The first method involves setting an arbitrary overall mastery level. The trainee either attains at least criterion or not. A second procedure is that of requiring all trainees to attain the same mastery level in a given objective but to vary the levels from objective to objective, depending on the difficulty of the material, importance of the method for later successful performance, etc. This second method seems more reflective of reality but as Prager et al. (1972) point out it is certainly more difficult to implement, let alone justify, specific levels that have been decided upon. Prager et al. believe that for handicapped children, at least, it would be appropriate to set mastery levels for each child relative to his potential. Nitko (1971) concurs and suggests different cut-offs for different individuals. However, the feasibility of individual cut-offs seems doubtful. Lyons (1972) points out that standards must take into account the varying criticality of the tasks. The criticality for any task is basically an assessment of the effect on an operating system of the incorrect performance on that task. Criticality must be determined during the task analysis and must be incorporated into the training objective. Unfortunately, in most cases the criticality of a task is not an absolute judgement and the selection of a metric for criticality becomes somewhat 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 high or very low proportion pass then we will obtain reliable measurement. Unfortunately, it is not often possible to "play around" with criterion scores to this extent. The training system may require a certain number passing and the criterion score is usually adjusted to provide the required number.

From this discussion it is apparent that there are no completely generalizable rules to guide the setting of cut-off scores. The cut-off must be realistic to allow the training system to provide a sufficient amount of trained manpower at some realistic level of competance. Training developers setting the cut-off score must therefore consider the abilities of the trainee population, the through-put requirements of the training system, the minimum competence requirement, and act accordingly. The use of summative try-out information should allow a realistic solution to the cut-off question for specific applications.

USES OF CRT IN NON-MILITARY EDUCATION SYSTEMS

Prager et al. (1972) describe research on one of the first CRT systems (Individual Achievement Monitoring System - IAMS) designed for the handicapped and designed for widespread implementation. Prager et al. point out that standardized tests often are useless when applied to handicapped individuals. They are simply too global in nature to be of much use in directing remediation. Tests build to reflect specific instructional objectives are much more useful when dealing with such populations. The use of CRTs also allows relating a handicapped child's progress to criterion tasks and competency levels. The use of CRTs 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 of these considerations, a CRT-centered accountability system has been devised. This project began with the construction of a bank of objectives and test items to mesh with the type of diagnostic individualization peculiar to the 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 decision: placement, immediate achievement, and retention. Standardized diagnostic and achievement tests were also used to aid in placement decision. The system is still in the early stages of implementation so no comment can be made concerning its ultimate usefulness.

More recently, Popham (1973) presents considerable data concerning 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 learners for a small period of time. At the conclusion of the "mini-lesson", the learners are given a post-test. Affective information is derived by asking the learners to rate the interest value of the lesson. Popham reviews 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. The teacher need not be the instructor, but the important aspect of this application

involves a post-lesson analysis in which the instructional approach is appraised in terms of its effects on learners.

3. A formative or summative evaluation device. Popham views this application of teacher performance tests to program evaluation to be extremely important, particularly in the appraisal of in-service and pre-service teacher education programs.

Popham presents three in-service and pre-service applications of the teacher performance tests. These applications were for the most part viewed as effective. However, a number of problems were revealed in the course of these applications that may be symptomatic of performance tests in general. Popham found that unless skilled supervisors were used in the conduct of the mini-lesson, most of the advantages of the post-lesson analysis were lost. Popham also found that visible dividends were gained by the use of supplemental normative information to give the teacher and the evaluation a bit more information regarding the adequacy of performance. In a similar area of endeavor, Baker (1973) reports the use of a teacher performance test as a dependent measure in the evaluation of instructional techniques. Baker discussed some shortcomings of the use of CRTs as dependent variables. These shortcomings are largely based on the peculiar psychometric properties of CRTs. However, Baker feels that CRT 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 the use of teacher performance tests with the indeterminate psychometric characteristics is not ethically permissible for evaluation of individuals--at least for the present.

In a slightly different area of application, Knipe (1973) summarizes the experience of the Grand Forks Learning System in which CRTs played a very salient part. The Grand Forks School District began by specifying in detail the performance objectives for K-12 in most subject areas. These objectives were to form the basis of a comprehensive set of teacher/learner contracts as one instructional method by which students could meet the objectives. It was found that mathematics was the subject area most amenable to analysis and therefore received the most extensive treatment. The mathematics test consisted of approximately 120 criterion-keyed items for each grade level 3-9. After extensive tryout the items were revised on the basis of teacher and student recommendations as well as on the basis of a psychometric analysis. The inclusion of psychometric analysis as a device to direct the revision of items seems questionable in view of the limited variance of CRTs. In summary, however, the teachers regarded the CRTs as useful in supplementing NRTs, and in addition found them useful for placement. Finally, Knipe concludes, "The criterion-reference test is the only type of test that a school district can use to determine if it is working toward its curriculum goals."

MILITARY USES

Extensive experience with use of CRT 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 in the administration of the 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 the training process. 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 & Smith present 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 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 among several criteria. Internal consistency reliability was also obtained. This effort is illustrative of good practice in CRT development and shows cleverness in the use of visual stimuli--the statistical treatments used in selecting items are, however, questionable. A somewhat similar development project entitled Learner Centered Instruction (LCI) (Pieper & Swezey), 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 seven volume document which includes a volume dealing with CRTs. A job performance orientation to CRT is advocated. Specific guidelines for task analysis and for translating criterion objectives into test items are presented in "hands-on performance" and in written contexts. The document is an excellent guide to the basic "do's" and "don'ts" in CRT construction. A similar Army document, TRADOC Regulation 350-100-1, Systems Engineering of Training presents guidelines for developing evaluation

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

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, if it is covered at all.

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 here. 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 on a number of Army installations has provided a diagrammatic guide to the development of instructional programs. CRT is not covered specifically in this document, nor is it addressed in the recent Army "state-of-the-art" report on instructional technology (Branson, Stone, Hannum, and Raynor, 1973). However, a CISTRAIN (Coordinated Instructional Systems Training) course (Deterline & 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 (Deterline and Lenn, 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.

U.S. Army Field Manual 21-6 (20 January 1967) provides trainers and instructors of U.S. Army in-service schools with guidance in the preparation of traditional instruction, e.g., lecturer, conferences, and demonstrations. FM 21-6 (20 January 1967) contains a great deal of information on construction of achievement tests but the "why's" and "how's" are largely lacking. The section on performance testing seems designed to discourage the construction and use of performance tests. In addition, the manual is weak on task analysis procedures -- procedures in general lack definition of method. All testing concepts are directed at the construction of norm-referenced tests of either job knowledge or performance. There is no discussion of how to set cut-offs, or any discussion of the issues peculiar to CRT. The emphasis is on relative achievement. Recently, FM 21-6 has undergone comprehensive revision to suit the needs of field trainers. The revised manual (1 December 1973) is generally in tune with contemporary training emphasis with considerable information on individualized training and team training. In particular, the extensive guidance provided on objective generation should prove very useful to field trainers. While the revised FM 21-6 does not specifically refer to CRT, the obvious emphasis on NRT which distinguished the 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 constraines of time and cost sometimes dictate modified standards, e.g., 80% 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.

From these limited examples it appears that the civilian sector has led in the development and use of CRTs. Although the EVATP effort is a notable exception, 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. School systems have had success with CRTs, largely due to the nature of the content domains chosen. These content domains heav y emphasize knowledge; hence tests can be paper and pencil which are cheap to administer. 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.

INDIRECT APPROACH TO CRITERION-REFERENCING

Fremer (1972) feels that is is meaningful to relate performance on Survey Achievement tests to significant real-life criteria, such as minimal competency, in a basic skills area. The author discusses various ways of relating survey test scores and criterion performance. All of these approaches are aimed at criterion-referenced interpretation of test scores. Fremer proposes 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. Fremer feels that considerable use can be made of the relationships observed among apparently diverse tasks within global content areas. Fremer further argues that tasks which are not samples of an objective may provide an adequate basis for generalization to that objective. Fremer notes 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 is offered of the use of 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 offers an illustrative example of using vocabulary test scores to define objective-referenced statements of ability to read editorials. Fremer notes, however, that the usefulness of interpretive tables, i.e., those that provide statements referencing criterion behaviors to a range of test scores, depends heavily on the method used to establish the relationship between the survey test scores and the objective-referenced ability. As 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 provides for the definition of several levels of mastery and points out that an absolute dichotomy, mastery versus non-mastery, will seldom be meaningful. It is difficult to understand why Fremer makes 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, Mastery/Non-Mastery. Fremer apparently bases 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. Fremer proposes that the notion of minimal competency should encompass a variety of behaviors of varying importance--the metric of importance will vary with the goals of the educational system.

Fremer (1972) proposes a method for relating survey test performance to a minimal competency standard that would involve a review of the proportion of students at some point in the curriculum who are rated as failures. This should serve as a rough estimate of the proportion of students failing to achieve minimal competency. It would then be 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 would be to acquire instructor judgement as to the extent to which individual items could be answered by students performing at a minimal level. By summing across items, it would be possible to obtain an estimate of the expected minimum score. Fremer, however, recognizes the limitations of this latter process with its high reliance on informed judgement. A further method proposed by Fremer seeks to define minimal competency in terms of student behaviors. The outcome of this method would be the identification of bands of test scores that would be associated with minimal competency. The processes involved in this method also rely on informed judgment, though.

Another method proposed by Fremer to criterion-referenced survey achievement tests involves developing new tests with a very narrow focus, i.e., a smaller area of content and a restricted range of difficulty. It should not be necessary to address every possible objective. However, it should be possible to develop a test composed of critical items by sampling from the pool of items. The next step in the process would involve relating achievement at various curriculum placements between the focused test and the survey instrument. This should allow keying of the items on the survey test with specific critical objectives.

Still another method put forth by Fremer to get from criterion-referenced to survey tests 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 that Fremer puts forth are very clever in concept and are mainly applicable to school systems and traditional curricula where well-developed survey instruments exist. Even so, considerable work is involved in keying the survey instrument. In non-school system instructional environments, dealing with non-traditional curricula, it is unlikely that an appropriate survey instrument would exist.

USING NRT TO DERIVE CRT DATA

Cox and Sterrett (1970) propose an interesting method for using NRTs to provide CRT information. The first step in this procedure is to specify curriculum objectives and to define pupil achievement with reference to these objectives. The second step would involve coding each standardized test item with reference to curriculum objectives. With coded test items and knowledge of the position of each pupil in the curriculum, it is possible to determine the item's validity in the sense that pupils should be able to correctly answer items that are coded to objectives that have already been covered. Step three is the scoring of the test independently for each pupil, taking into account his position in the curriculum. The authors recommend 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 pupil 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. This method seems an economical way to extract CRT information and NRT information from the same instrument. The technique has yet to be explored in practice, however.

CONSIDERATIONS FOR A CRT IMPLEMENTATION MODEL

The development and use of CRT is a fairly recent development in instructional technology. Partially as a result of this, there is no comprehensive theory of CRT such as exists for NRT. Hence, the concepts of validity and reliability for CRT are not yet well developed, although definition of these concepts is necessary to reduce errors of classification. The need for content validity in CRT is, however, well recognized. In addition, there is no single CRT construction methodology which will serve for all content

domains. Unresolved questions also revolve around the question of Bandwith fidelity and the use of reduced fidelity in criterion-referenced performance tests.

The rationale for the use of CRT in evaluating training programs and describing individual performance is well established. To ensure best possible results, the military or industrial user 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 knowledge that must be trained.
    c. Careful identification of job conditions
    d. Careful identifcation of job standards
    e. Identification of critical tasks.
2. Careful formulation of objectives
    a. Particular care in the setting of standards
    b. Identification of all enabling objectives
    c. Independent check 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. Determination of item form
    d. Statement of rules for items
    e. Development of item pool for objectives to be tested
    f. Develop tryout plan and criteria for item acceptance
    g. Tryout of items
    h. Revision and rejection of items.

Particular care must be exercised in setting item acceptance criteria for item tryout. The use of typical NRT item statistics should be minimized. The usual methods are totally inadequate, i.e., internal consistency estimates are only suitable with large numbers of items; in addition, internal consistency may not be an important consideration. Traditional stability indexes may also be inappropriate due again to small numbers of items and reduced variance. The technique proposed by Edmonston et al. (1972) may prove effective in reducing errors of misclassification due to inadequate test items.

By adhering to strict quality control measures, it should be possible to obtain a set of measures that have a strong connection with a specified content domain. Whether or not they are sensitive to instruction, or if they 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 CONSIDERATION

Although the costs of training and the costs of test administration can readily be quantified in dollar terms, we lack a proper metric to completely assess the costs of misclassification. Emrick (1971) proposes 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 the task analysis. If the analysis of the job reveals a large number of critical tasks or individual tasks whose criticality is great, then the cost of supplying a non-master can be assessed as high, and great effort is justified in developing a training program featuring high fidelity, costly CRT. Where the analysis does not reveal high numbers of critical tasks, the cost then becomes a function of less quantifiable aspects. 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 would result in greatly reduced testing costs in many instances. The decision as to the appropriateness of the test must be made empirically on the basis of well controlled tryout with typical course entrants. The development of symbolic performance tests may prove to be difficult. Much is yet to be known about how to approach this development. If progress can be made in lowering the cost of CRT then the problem of cost-benefit analysis will be made in lowering the cost of CRT then the problem of cost-benefit analysis will be largely obviated.

As the question currently stands, there is no doubt that CRT 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 CRT is a must.

PART 2--SURVEY OF CRITERION-REFERENCED TESTING IN THE ARMY

PURPOSE AND METHOD OF THE SURVEY

In order to survey the application of criterion-referenced testing techniques in the military, a number of Army installations were visited. Information was collected to supplement the literature search and review, to provide detailed material on CRT development and use in the Army, and to obtain information on attitudes and opinions of Army testing personnel.

Specifically, the survey gathered data on:

1. How CRTs are developed for Army applications. In order to create a CRT construction manual which will be useful to Army test developers, it is necessary to determine how CRTs are currently developed in the Army. Additionally, it is important to determine differences in test development strategies across Army installations, so that the manual can suggest procedures which will mate well with a variety of approaches.

2. How CRTs are administered in various Army contexts. This information is important since design for administration materially affects the test construction process. Design information is important in creating guidelines on development of CRTs, in order to make them suitable for administration in diverse, Army testing situations.

3. How CRT results are used in the Army. The way in which a test's results are used is a factor that must be considered in the development of any test. Hence, the survey obtained data on use of test results in a variety of Army testing situations.

4. Extent of criterion-referenced testing in the Army. This includes information on extensity--how prevalent criterion-referenced testing is in the large, Army-wide sense; and information on intensity--how much testing in specific Army contexts is of a criterion-referenced type.

5. The level of personnel who will use the CRT Construction Manual developed by the project. This information includes educational levels, range of military experience, and familiarity with psychometric concepts. Such information is designed to help tailor the manual to its audience.

6. Problems encountered by Army testing personnel in the development and use of criterion-referenced tests. Information on problems serves two purposes. First, the identification of typical problem areas points the way toward future research on criterion-referenced testing. Second, the CRT Construction Manual can deal with typical problems, offering suggestions for avoiding or surmounting them.

7. Attitudes of Army testing personnel toward the development and use of CRTs. It is important to assess existing attitudes toward CRTs among Army testing personnel, since level of acceptance is an indicator of spread and utility of a new concept. Additionally, attitudinal data will enable the CRT Construction Manual to address current attitudes, and thus to attempt to rectify poor attitudes based upon misconceptions.

8. The probable future course of criterion-referenced testing in the Army. Interview data, particularly that collected from personnel at supervisory levels, indicate probable trends in future Army CRT use. Also, problems in implementing CRT applications suggest needed research.

9. Sample Army CRTs and problems in developing and using them. An important part of the on-site survey is to gather materials to serve as the basis for examples of CRT development and use.

Interview Protocol Development. In order to gather these types of information, an interview protocol for on-site use at various Army posts was developed. Development of the protocol included several review phases during which revised versions of the protocol were prepared. The second version of the protocol consisted of three forms: One to be used in interviews with test constructors, another for test users, and a third to be used with supervisory personnel. The final instrument combined these forms and included several optional items for use in interviews with personnel who were especially knowledgeable about criterion-referenced testing. The final version of the protocol was found to have high utility, since it can be used to structure interviews with personnel who serve any of three functions (test construction, test use, and supervision). The protocol provides flexibility in the range of topics to be discussed in an interview, thereby allowing interviews to be tailored to the ranges of responsibilities, experience, and knowledge possessed by individual interviewees. Appendix A of this report is a copy of the final version of the protocol.

The interview protocol was used in a series of one-to-one interviews conducted during January, February and March 1974. Installations surveyed during this period included the Infantry School at Fort Benning, the Artillery School at Fort Sill, the Air Defense School at Fort Bliss, the Armor School at Fort Knox, and BCT and AIT units at Fort Ord. In addition, test-related departments were surveyed at each post. A total of 105 individuals were interviewed.

Survey Teams. A survey team spent three days at each post surveyed. The interviews ranged in duration from approximately one-half to three hours apiece and averaged about one and one-half hours. Interview length was at the interviewer's discretion, based on the utility of the information obtained from a subject.

Summaries of the types of personnel surveyed at each installation, presented in a following section of this report, indicate each interviewee's position in the organization for Army School, MOS, TEC, and Training Center testing programs, and whether the individual is a test developer/user (test administrator, test scorer, etc.) or a supervisor of test construction or use.

Each interviewee responded to most of the items on the protocol. Responses which are easily and meaningfully quantifiable are presented in tables in the following section. Other items elicited opinions, anecdotal information, process information, and other data that are not easily quantifiable. Such data are summarized by extracting and comparing verbal descriptions and are also discussed in the next section of this report.

RESULTS AND DISCUSSION

Sample. Table 1 presents a summary of the individuals interviewed at Forts Benning, Bliss, Sill, Knox, and Ord. Of 105 individuals interviewed, more than half were personally involved in constructing, administering, scoring, or making decisions based on test scores. The remaining individuals surveyed were supervisors of personnel who constructed or used tests.[1]

Table 1 also identifies four categories of subjects: School personnel (Infantry, Artillery, Air Defense, and Armor), Military Occupational Specialty (MOS) Test personnel (groups involved with the development and administration of annual MOS tests), Training Center personnel (BCT and AIT),

---

[1] Also included in the survey was a visit to the U.S. Army Southeastern Signal School (USASESS) and the U.S. Army Military Police School, Fort Gordon, Georgia. Contractual time constraints did not allow the application of the formal survey protocol. Following is a summary of the findings at Fort Gordon.

Test Quality Control at USASESS is conducted on both an internal and external basis. Internal control entails examining tests constructed by the academic departments for consistency using Evaluation Planning Information Sheets (EPISs). These documents are in turn, examined for consistency with Training Analysis Information Sheets from which they are derived. Examinations are supplemented by direct observation of on-going tests, to ensure that requirements in the test administrator's manual are being met (i.e., that appropriate tasks, conditions, and standards are being employed).

External quality control is maintained through the use of questionnaires which ask field unit respondents to indicate the actual job value of tasks on which they were trained in school. The questionnaires are followed up by direct interviews with school graduates in the field. Additional quality control information is obtained via communication with field unit commanders.

Among the problems noted were: (1) concern for lack of adequate criteria in training the "soft skills", such as counseling and leadership; and (2) ambiguity in existing regulations are open to varying interpretations by different schools and by individuals within schools.

and TEC (Training Extension Course) Program personnel. No Training Center data were collected at Fort Benning, while Fort Ord data were exclusively with Training Center programs.

A total of 67 individuals were interviewed in School organizations. This focus on school personnel is appropriate since the CRT Construction Manual will be used primarily in the schools. It is interesting to note that of the 79 subjects who were asked if special training were available for testing personnel, almost 80% responded yes. This does not mean that 80% of the subjects asked had received such training, but that training in testing techniques is available in the Army. Many individuals who participated in the survey were experienced in constructing or administering tests, and several had received special training in testing. For a more detailed analysis of the subjects and their organizational positions, see Appendix B.

Tables 2 through 7 present summaries of responses to quantifiable protocol items. The data upon which these summaries are based are in Appendix C. Note that since interviews were tailored to address the knowledge and experience of the individual, not all subjects were asked all items. For example if it was established that an individual was not involved in test development but in test administration or in use of test results, that individual was not queried concerning test construction. Hence, in Table 2, for example, a maximum of 87 individuals responded to a given item.

Test Development. Table 2 summarizes responses to protocol items concerning involvement with various steps of CRT development. Details of Army test construction processes vary widely; however, some impressions of the test construction process can be gained from Table 2.

The data presented in Table 2 are subject to interpretation. For example, although slightly over half of the 50 subjects answered "yes" to the protocol item about using an item analysis technique (item 9b), further questioning during the interview usually revealed that they were not using a formal item analysis technique. Instead, they typically inspect a computer printout of percent right and wrong responses to items on a test. Items having an unusually high number of wrong responses are reworked or discarded.

After the final test items are selected, Army test developers usually do not assess reliability and validity, at least in a strict psychometric sense. Instead, the tests are administered several times and items that cause a great deal of difficulty are reviewed to see if they are constructed properly--a relatively informal process.

Table 1

SURVEY OF CRITERION-REFERENCED TESTING IN THE ARMY: SUBJECTS INTERVIEWED AT FORT BENNING, FORT BLISS, FORT SILL, FORT KNOX, AND FORT ORD
(N = 105[a])

		School	MOS	Training Center	TEC Program
Ft. Benning, Georgia	S	5	1	0	1
	TDU	14	1	0	1
Ft. Bliss, Texas	S	7	0	0	1
	TDU	12	0	3	2
Fort Sill, Oklahoma	S	7	1	0	2
	TDU	8	2	0	1
Ft. Knox, Kentucky	S	7	1	1	0
	TDU	7	0	0	0
Ft. Ord, California	S	0	0	10	0
	TDU	0	0	10	0
Totals		67	6	24	9

[a] Total Number of Supervisors (S) Interviewed: 44

Total Number of Test Developers/Users (TDU) Interviewed: 61

Table 2

INVOLVEMENT IN VARIOUS STEPS OF TEST DEVELOPMENT:
SUMMARY OF RESPONSES ACROSS ALL POSTS

Item No.	Brief Statement of Item[a]	Number of Subjects Responding to Item	Percent of "Yes" Responses
4	Have you been included in writing objectives	76	78
4b	Do you write objectives in operational, behavioral terms?	42	71
5	Have you participated in setting standards?	69	77
6	Have you participated in imposing practical constraints?	72	68
7	Have you helped determine priorities?	70	67
8	Have you been included in writing test items?	68	70
8b	Do you write item pools?	50	66
9	Have you been involved in selecting final test items?	67	[illegible]
9b	Do you use an item analysis technique?	50	[illegible]
11	Do you measure test reliability?	[illegible]	33
11b	Do you compute coefficients of reliability?	42	[illegible]
12	Do you aid in validating tests?	[illegible]	35
12b	Do you use content validity?	41	[illegible]

[a] For complete wording of the protocol items, see Appendix A

It appears that relative care is taken in Army test development programs to select and define objectives and their associated conditions and standards. Some care is taken in writing items to match these objectives. From this point on, however, empirical rigor is lacking; that is, formal item analysis and assessment of test reliability and validity are infrequently done.

Test Administration. Table 3 presents subject responses to protocol items dealing with test administration. A large proportion of subjects in the survey have been involved in administering tests. This is not surprising since much test development is done by school instructors; thus, individuals who create test items also administer the tests in their classes. These are heartening data: It is advantageous for test developers to be familiar with test administration situations, since it gives them increased familiarity with the conditions and limitations inherent in such situations.

Table 3

INVOLVEMENT IN ASPECTS OF TEST ADMINISTRATION:
SUMMARY OF RESPONSES ACROSS ALL POSTS

Item No.	Brief Statement of Item[a]	Number of Subjects Responding to Item	Percent of "Yes" Responses
10	Have you participated in administering tests?	85	86
10b	Do you ever use the "assist method"?	75	69
13	Do you use "go-no go" scoring standards?	100	49
14b	Do you retest trainees who fail the first time?	55	71

[a] For complete wording of the protocol items, see Appendix A

Table 3 also shows that an "assist" method of scoring is frequently used. It appears that test administrators often find it appropriate to provide help to individuals taking the test. The actual percentage of test administrators using a true assist method is probably somewhat lower than that shown in Table 3, since a good number of those who stated that they use this method indicated that they provide help only if testees have difficulty with ambiguities in test language or instructions. In a true assist method, help is given to those individuals who can not perform a particular item for whatever reason. Such a method is often used in cases where the testee could not otherwise complete the test (e.g., a checkout procedure).

Less than half of the 100 subjects queried said that they used go-no go scoring standards on their tests. This does not imply that more than half of the individuals in our survey necessarily use normative scoring standards; instead, many use point scales for scoring.

Over 70% responded that trainees who fail a test the first time are retested. There are many cases where retesting is done. For example, in BCT, AIT and other hands-on performance testing situations, trainees are often given second and third chances to pass particular performance items.

Uses of Test Results. The primary use of test results is, of course, to evaluate individual performance. This is true whether the test is criterion-referenced or normatively based. There are, however, other ways in which test results can be used. Table 4 presents a summary of responses to protocol items dealing with various uses of test results. Table 4 shows that the most common uses of test results, other than for evaluation of trainee performance, are for improving training and for diagnosis. Test results can diagnose areas in which an individual is weak and in need of remediation. Seventy-two percent of the subjects questioned indicated that they use test results for diagnostic purposes. Diagnosis is usually done informally: Instructors review test results and then confer with trainees.

Test results can also be used to assess course adequacy in the formative evaluation sense. Seventy-three percent of the subjects questioned indicated that they use feedback from the tests to improve courses. The way in which this feedback is used varies widely. For example, some senior instructors indicated that if many trainees from a particular instructor's class perform poorly on certain parts of a test, they would first evaluate the instructor. If several classes taught by different instructors scored poorly on a section of a test, the senior instructor might review the materials used in that portion of the course. In other situations, the test itself is reviewed using feedback from the students. For example, if a test item is unclearly worded or if the performance called for is unclear, student feedback is a valuable tool.

Table 4

USE OF TEST RESULTS OTHER THAN EVALUATING INDIVIDUAL PERFORMANCE: SUMMARY OF RESPONSES ACROSS ALL POSTS

Item No.	Brief Statement of Item[a]	Number of Subjects Responding to Item	Percent of "Yes" Responses
14	Do you use test results to compare trainees?	91	63
15	Do you use test feedback to improve courses?	96	73
16	Do you use test results for diagnostic purposes?	93	72
28	Are you familiar with team performance testing?	88	42

[a] For complete wording of the protocol items, see Appendix A

Less than two-thirds of the subjects questioned indicated that test results are used to compare trainees. Comparing individuals on the basis of test results is essentially norm-referenced. It is possible however, to employ CRTs for norm-referenced purposes. In BCT, for instance, trainees who pass the comprehensive performance test on their first try might be considered for promotion from E1 to E2, while those who do not may not be so considered.

Considerably less than half of the subjects questioned said that they were familiar with team performance testing situations. Further, of those who indicated familiarity with the concept, many indicated that team performance testing is often individual evaluation in a team context. Actually, the testing of team performance was very limited on the Army posts visited.

Types of Tests. Table 5 shows a description of types of tests constructed or used by subjects in our survey sample, based upon their responses to protocol item 27. Part 1 is a categorization according to test mode, Part 2 according to test use. For both parts, subjects were asked to indicate the approximate percentage of each type test with which they were involved.

Table 5

TYPES OF TESTS CONSTRUCTED OR USED: SUMMARY OF RESPONSES TO PROTOCOL ITEM 27 ACROSS ALL POSTS

Item 27 - Part 1
N = 93

What proportion of the tests you have participated in making or using are:

	Mean Response
A. Paper-and-pencil knowledge tests?	47.2%
B. Simulated performance tests? (e.g., using mockups and drawings)	7.9%
C. "Hands-on" performance tests?	41.1%
D. Other?	3.8%
Total:	100%

Item 27 - Part 2
N = 75

What proportion of the tests you have participated in making or using are for:

	Mean Response
A. Specific skill and knowledge requirements?	39.4%
B. Specialty areas in a course?	7.3%
C. End of block within a course?	30.0%
D. Mid cycle within a course?	5.3%
E. End of course?	16.0%
Total:	100%

It appears that most tests are either paper-and-pencil knowledge tests or hands-on performance tests. Although Table 5 indicates that paper-and-pencil knowledge tests are nearly 50% of those created and used, many subjects confused paper-and-pencil knowledge tests with paper-and-pencil performance tests. This was learned from discussions with interviewees. In many areas, paper-and-pencil tests are equivalent to the performance called for in the actual task situation. For example, such diverse areas as map-making and aiming artillery require paper-and-pencil performance. Maps must be drawn to scale, while in many cases the aiming of artillery requires mathematical computations. It is estimated that about half of the responses in the paper-and-pencil knowledge test category actually referred to paper-and-pencil performance testing. Thus, responses to Part 1 of Item 27 can be interpreted to indicate that nearly three-quarters of the tests constructed or used are performance tests of one sort or another. These results accord with the emphasis on performance testing, and indicate that performance testing has become widespread in many phases of Army evaluation.

Responses to Part 2 indicate that tests measuring specific skill and knowledge requirements, and those used at ends of blocks of instruction, account for about 70% of test construction and use. Mid-cycle tests and end-of-course tests together account for less than one-quarter of the tests. Responses to Part 2 of Item 27 indicate that tests are well distributed throughout instruction. This is good news since frequent testing can provide frequent feedback and the possibility for on-going remediation.

Problems. Table 6 presents a summary of responses to protocol items dealing with problems in the development and use of CRTs. Over two-thirds of the subjects (who were primarily supervisory personnel for this item) indicated that increased expense may be a problem in the development and use of CRTs. Several subjects commented that the extra expense may be a factor in reducing the availability of CRTs in the Army. However, many individuals indicated that increased expense is a short-term factor, and that in the long run, criterion-referenced testing is less expensive than is norm-referenced testing. Criterion-referenced testing is presumably less costly in terms of insuring the efficient output of well-trained soldiers.

Many individuals in the survey sample felt that time pressures, or other constraints, often prevent successful construction and use of tests. In discussion, subjects indicated that time pressure is the most common constraint, and that time pressures are usually present in test development.

Table 6

GENERAL PROBLEMS IN THE DEVELOPMENT AND USE OF CRITERION-REFERENCED TESTS: SUMMARY OF RESPONSES ACROSS ALL POSTS

Item No.	Brief Statement of Item[a]	Number of Subjects Responding to Item	Percent of "Yes" Responses
30	Have time pressures, or other constraints prevented successful test, test construction and use?	89	61
31	Have you seen tests which were unsuitable for their intended uses?	54	37
35	Are Criterion-Referenced Tests more expensive to develop and use than norm-referenced tests?	49	71

[a] For complete wording of the protocol items, see Appendix A

However, time pressures and other constraints do not usually interfere with test administration tasks. Usually, tests are administered satisfactorily despite time pressures. Interviewees seemed to think that Army test development and administration have improved greatly in recent years.

Attitudes. Table 7 presents a summary of subject attitudes concerning criterion-referenced testing in the Army. In general, subjects were in favor of the Army trend toward criterion-referenced testing. Comments included: "Criterion-referenced testing is the best system of testing yet devised"; "It is the only way to go"; "It is a terrific improvement over testing in the old Army"; "Criterion-referenced testing should be used exclusively in the Army and wherever else possible, including civilian educational institutions." Eighty-eight percent of the individuals responding felt that criterion-referenced testing should receive high or top priority in terms of Army assessment programs. Sixty percent felt that criterion-referenced tests should replace most or all norm-referenced tests.

Subjects felt that criterion-referenced testing is practical and useful in measuring job performance skills. No other item on the survey protocol elicited a 100% positive response. In addition, many individuals felt that criterion-referenced testing would be useful and practical for measuring

Table 7

ATTITUDES CONCERNING CRITERION-REFERENCED TESTING: SUMMARY OF RESPONSES TO PROTOCOL ITEMS 34 AND 40 ACROSS ALL POSTS

Item 34

How strongly do you feel about future use of Criterion-Referenced Testing in the Army? Should Criterion-Referenced Test development receive high or low priority in terms of Army assessment programs?

N = 80

Percent Responding to Each Alternative	
1	Strongly against--Criterion-Referenced Testing should receive bottom priority, or dropped entirely.
1	Against-Criterion-Referenced Testing should receive low priority.
10	Neutral--Criterion-Referenced Testing should receive average priority.
28	For--Criterion-Referenced Testing should receive high priority.
60	Strongly for--Criterion-Referenced Testing should receive top priority, Criterion-Referenced Tests should replace most or all norm-referenced tests.

Total: 100%

Item 40

Do you feel that Criterion-Referenced Testing is practical and useful in measuring job performance skills?

Number of Interviewees Responding = 64

Percent responding "yes" = 100

areas other than job performance skills. Knowledge tests, for example, were seen by many as a practical and useful application of the criterion-referenced concept.

DISCUSSION OF CRT SURVEY

Over 150 hours of interviews were conducted during the survey of criterion-referenced testing in the Army. Topics covered ranged from the extent, utility, and practicality of CRT use in the Army, to problems in implementing CRTs.

Although criterion-referenced testing is used in today's Army, many NRTs are in use also. This is not surprising, since criterion-referenced testing is a relatively new concept. It was apparent from the survey, however, that CRT use is increasing.

At each installation visited, criterion-referenced testing was in evidence. The combat arms schools visited--Infantry, Armor, Artillery and Air Defense--develop and use a number of CRTs. However, school implementation of criterion-referenced testing is in the beginning stages. Some departments are making serious attempts to incorporate CRTs, while others are only minimally involved. Many employ criterion-referenced terminology, but do not produce true CRTs. This is especially true in "soft skill" areas, such as tactics and leadership. Most academic departments within these four combat arms schools indicated that many of their tests, especially the written ones, are graded on a curve. Much reliance appears to be placed upon subjectively graded paper-and-pencil tests and upon computer-graded objective tests.

MOS testing continues to be primarily norm-referenced. Most, if not all, MOS tests rely on situational multiple-choice items. Because of the low fidelity of such items, it is often difficult to determine if they are criterion- or norm-referenced. On the surface, at least, they are suspiciously similar to conventional knowledge test questions.

Consideration of the CRT concept is being given to Training Extension Course packages. The optional "audio-only" performance test appended to such TEC packages requires further development and implementation so that TEC instruction can be more thoroughly evaluated in a criterion-referenced fashion.

At Fort Ord, California, CRTs are employed both in BCT and in AIT. Although there are problems in the administration of the Comprehensive Performance Tests (a type of CRT used toward the end of basic training) the testing experiences at Fort Ord should be able to serve as a good "field laboratory" for developing CRT applications.

AIT in diverse areas such as field wiring and food services appears to be benefiting from the use of CRTs. Preliminary indications are that more soldiers are being evaluated more effectively through the application of criterion-referenced testing. Further, instructors, supervisors and students all appear to be favorably disposed toward CRTs.

In general, although criterion-referenced testing is not extensive, there are many instances of serious attempts at CRT development and use at the Army installations visited.[2] There was much respect for the utility and practicality of criterion-referenced testing. As noted, many interviewees were strongly in favor of increased use of criterion-referenced testing in the Army. Many who had experience with developing or using such tests indicated increased evaluation effectiveness, increased individual morale and, in the long run, reduced expense as a function of CRTs. Despite this high regard, there was too little rigorous development or application of CRTs. While progress is being made toward achieving rigor in "hard skill" areas, especially in equipment-related skills, attempts in "soft skill" areas are lacking. Personnel who develop tests for such areas in many cases are attempting to develop CRTs, but are diverted at the outset since genuine difficulties in specifying objectives explicitly are often encountered.

The survey revealed virtually no evidence of criterion-referenced testing in team performance situations. In fact, as many subjects pointed out, operational units are not formed until after AIT. This does not mean, however, that CRT development for unit performance is inappropriate. Such tests could be developed and used in AIT and then exported to field units. Although problems may occur when an individual begins to work within a field unit, this is not an argument against unit CRTs.

The CRT Construction Manual. Subjects at all levels indicated a need for increased development and use of criterion-referenced testing in the Army. Many indicated the need for guidance in constructing and in administering CRTs. A consensus indicated that such guidance should be written in simple, straightforward language and should address criterion-referenced testing in a non-theoretical, practical manner. Individuals interviewed in the survey indicated that a manual of this type would be well received at all levels in test development and evaluation units.

---

[2] Many of the personnel interviewed confused CRTs with "hands-on" performance testing. In terms of implementing hands-on performance testing programs, the trend at the Army posts visited is dramatic; many such tests are in evidence. Not all of these tests are criterion-referenced, however; many are not. In order to be called criterion-referenced, an individual testee's skills or knowledges must be compared to some external standard. This means that test items must be matched to objectives which are derived from valid performance data. This is not the case for a significant proportion of the "hands-on" performance tests presently used at the sites surveyed.

Test Development Process. A number of difficulties in CRT development and use were observed and/or described during the survey. First, the development of CRTs must be derived from well-specified objectives which are, in turn, the results of careful task analyses. Unfortunately, task analysis data are not available in many cases, and in cases where they are available, they are often disregarded. Many test developers write statements of performance standards from Plans of Instruction (POIs) or from Army Subject Schedules. In most cases, these POIs and schedules are based upon task analyses. However, often the critical source data are not readily apparent. In other cases, objectives are defined "out of the blue" by subject matter experts who may be unfamiliar with the instructional system development process. Worse yet, in some cases careful task analyses have been developed and then ignored. For example, in one AIT course visited, a careful task analysis had been conducted which accurately documented critical behaviors. Although the performance tests used in the course were developed from objectives derived from the task analysis, the recently revised subject schedule ignored, and in some cases flatly contradicted, the task analytic data. As a result, the revised subject schedule required testing skills that the task analysis had revealed are performed very infrequently; but did not mention other skills which, according to the task analysis, were most frequently performed.

Many difficulties in CRT development can be overcome if task-analytic data are actually used in the development of tests. When tests are modified for local administration, those responsible for the modification should have access to the same task-analysis data.

Practical Constraints. The CRT survey suggested that priorities and practical constraints for task objectives are usually assessed informally. If task priorities are not accurately assessed and defined, the development of test items which measure the achievement of objectives is exceedingly difficult. If all objectives are taken to be of equal weight, then they will normally be assessed by an equal number of test items when, in fact, more important objectives may require more thorough testing.

Frequently, practical constraints to the testing situation are considered only as an afterthought. Constraints which operate in the testing situation should rightfully be considered while a test is being developed. Some Soldier's Manual Army Testing (SMART) books, for example, show a minimal regard for practical testing constraints. They contain lengthy checklists which, although possibly of use in evaluating an individual's performance, cannot be followed by test administrators. In some cases, one tester may administer a SMART test to many soldiers simultaneously, although totally unable to observe all items on the SMART checklist. Thus, at a given testing station, a particular soldier may be scored as a "no go" while another soldier may be scored "go" because the tester could only observe one accurately. The problem of including practical testing constraints and task priorities can be solved by training test developers to consider these as an integral part of the test development process.

Item Pools. Test developers seem to have little difficulty creating items if the performances, standards, and conditions are accurately specified. However, many Army test developers surveyed indicated that they wrote only the precise number of items required for a specific test. These items are typically reviewed by subject matter experts and are then revised accordingly. If alternate forms of a test are required, a pool of items are constructed such that a computer can format alternate test forms by selecting a subset of items from the pool. Rarely are extra items written. Accordingly, there is no empirical selection process for final test items. Items are typically dropped or revised, after a review, if large numbers of individuals in a class answer them incorrectly.

Creating a test item pool should become a standard part of the test development process. If twice as many items are developed as are needed for a specific test, the test can be tried out and the final items selected empirically. An empirical item analysis strategy should be incorporated to select final test items. Although the creation of item pools and the use of item analysis techniques may introduce added expense into the test development procedure, the payoff should outweigh the expense. The payoff here is the development of items that are feasible and which reliably address appropriate criterion behaviors.

Reliability and Validity. A major omission in the development of CRTs, as observed during the Army survey, is the lack of test evaluation. There was virtually no consideration of test reliability and/or validity. This does not indicate that the tests as developed are unreliable, but that the question has not been addressed. A few subjects did indicate that content validity had been considered by virtue of careful matching of test items and task objectives. Content validity however, is not necessarily the only type of validity appropriate for CRTs. Predictive validity can also be assessed. That is, trainees can be tested using CRTs and then evaluated under field conditions performing the tasks for which they have been trained. Test results for a valid test should be congruent with later field performance results.

Army test developers should be instructed in techniques for establishing reliability and validity of CRTs. Even if a test evidences content validity as a function of careful creation based upon task objectives, reliability is still in question. If a test cannot be administered reliably, results are meaningless.

Administration. A poorly administered test defeats long hours of careful test development. The CRT survey indicated that lack of standardized testing conditions exist in many areas. This is in part attributable to lack of training in test administration for testers, and in part to lack of clearly defined test administration instructions.

One administrative problem observed was that soldiers may be aided or hindered as a function of their position in the performance testing line. Those who are not first in line "get a break" by observing mistakes of others.

The test administrative conditions should specify that trainees waiting to be tested remain at a certain distance from the test site, or the test administrators should be instructed in conducting such tests in standardized manner, or both.

Careful instructions in test administration are necessary to insure accurate testing. Steps should be taken to insure that test administration practices are clearly defined for each test, and that test administrators are adequately trained. Further, test sites should be regularly inspected to insure that tests are being given under the specified standard conditions.

REFERENCES

Ammerman, H. L., and 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 square estimates of item parameters for the mastery learning test model: Fixed GMP version. SWRL, 1973. (a)

Besel, R. R. Program for computing least squares estimates of item parameters for tne 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.

Branson, R. K., Stone, J. H., Hannum, W. H., and 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.

Chenzoff, A. P. A Review of the Literature on Task Analysis Methods. Valencia, PA: Applied Science Associates, Inc. (Tech. Rep. 1218-3), 1964.

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

Cox, R. C. and 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. and Gleser, G. C. Psychological tests and personnel decisions. Urbana: University of Illinois Press, 1965.

Cronbach, L. J. and 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.

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

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

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., and 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., and Rehder, R. J. A comparison of correlated-job and work-sample measures for general vehicle repairmen. HumRRO Technical Report 70-16, 1970.

Flanagan, J. C. Critical requirements: A new approach to employee evaluation. Personnel Psychology, 1949, 2, 419-429.

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

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, and 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., and 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., and Kruskal, W. H. Measure of association for cross classification. American Statistical Association Journal, 1954, 49, 732, 764.

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

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

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

Hambleton, R. K., Rovinelli, R., and Gorth, W. P. Efficiency of various item-examinee sampling designs for estimating test parameters. Proceedings, 79th Annual Convention of American Psychological Association, 1971.

Hanson, R. A., and 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 test designed to measure achievement following performance-oriented instruction. Paper presented at the Annual Meeting of the American Psychological Association, Honolulu, 1972.

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

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

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

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.

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. University of Pittsburgh Press, 1964.

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

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., and 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., and 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.) Psychological principles in system development. New York: Holt, Rinehart, and Winston, 1962.

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.

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.

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., Catrow, E. J., Swezey, R. W., and 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.

Pieper, W. J., and 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.

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.

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

Prager, B. B., Mann, L., Burger, R. M., and 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.

Rahmlow, H. R., Matthews, J. J., and 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., and Summer, G. Some considerations in the experimental design and evaluation of educational innovations. Santa Monica, Calif: Rand Corporation, 1970.

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. and 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.

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

Smith, R. G., Jr. The development of training objectives. HumRRO Research Bulletin 11, 1964.

Swezey, R. W. Criterion-referenced measurement in job performance assessment. Reston, Va: Applied Science Associates, 1974.

Taylor, J. E., Michaels, E. R., and 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. Training handbook. Fort Benning, Ga.: U.S. Army Infantry School.

Tyler, Ralph W. Constructing achievement tests. Columbus: Ohio State University, 1934.

Tyler, Ralph W. Basic principles of curriculum and instruction. University of Chicago Press, 1950.

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

Vineberg, R., and 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.

ADDITIONAL REFERENCES

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

U.S. Army. How to prepare and conduct military training. Field Manual 21-6. Washington, D.C. 20 January 1967.

U.S. Army. Systems engineering of unit training. CONARC Pamphlet 350-11. Fort Monroe, Va.: Headquarters, U.S. Continental Army Command (now U.S. Army Training and Doctrine Command), 12 January 1973.

U.S. Army. How to prepare and conduct military training. Field Manual 21-6, Washington, D.C. 1 December 1973.

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

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

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

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

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

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

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).

APPENDIXES

APPENDIX A

INTERVIEW PROTOCOL:
SURVEY OF CRITERION-REFERENCED TESTING IN THE ARMY

* = Optional question: Ask as appropriate

Name of Interviewee: _____

Mailing Address: _____

_____

_____

Telephone Number: _____

Introduction. Interviewer will:

A. Introduce himself

B. Introduce ASA

C. Explain that ASA is doing contract work for the Army Research Institute

D. State that ASA is interested in improving tests for the Army

E. Explain that ASA wants to find out about current status of testing in the Army so we can determine what we can build on

1. What is your position in the organization here?

   What school or center are you in? _____

   What is your directorate, department, or unit? _____

   What is your branch or section? _____

   What is your position and title? _____

_____

_____

_____

2. How long have you been involved in testing? Years_____ Months_____

3. What did you do before you became involved in testing?

_____

_____

_____

Interviewer Statement: Now, I would like to discuss with you, some tasks that may be involved in test construction and use. These tasks are done in different ways in different places. Sometimes they are combined, in other cases some are eliminated. They often go by different names. Would you please tell me which of these you are involved in.

* 4. Writing objectives. That is--determining what the test will measure and the conditions under which the measurement will occur in terms of precise, behavioral statements.

Have you been involved in writing objectives? Yes_____ No_____

If yes, (a) how long have you been doing this? Years_____ Months_____

(b) do you write objectives in operational, behavioral terms?

Yes_____ No_____ Don't understand_____

* 5. Setting standards. That is--defining the standards against which performance is evaluated. In many cases, these standards are very similar to the stated objectives.

Have you participated in setting standards? Yes_____ No_____

If yes, how long have you been doing this? Years_____ Months_____

* 6. Imposing practical constraints. That is--deciding how the test must be built so it can actually be used within the limits of the situation for which it is designed. For example, there are often time constraints involved in testing complex skills.

Have you been involved in this? Yes_____ No_____

If yes, how long have you been doing this? Years_____ Months_____

* 7. Determining priorities. That is--deciding how important each standard is in relation to other standards.

Have you helped determine priorities? Yes_____ No_____

If yes, how long have you been involved in determining priorities?

Years_____ Months_____

* 8. Writing items. That is--creating items for use in the test.

Have you written, or helped to write items? Yes_____ No_____

If yes, (a) how long have you been involved in writing items?

Years_____ Months_____

(b) does your group of items usually contain more than will be included in the test? Yes_____ No_____ Don't know_____

* 9. **Selecting final test items. That is--applying statistical tests to determine the most useful, non-redundant items.**

Have you been involved in selecting final test items? Yes_____ No_____

If yes, (a) for how long have you done such work? Years_____ Months____

(b) do you use an item analysis technique?

Yes_____ No_____ Don't know_____

* 10. Test administration. That is--administering the test in the situations for which it was planned. Also, test administration is often done as a try-out, before the test is finalized.

Have you participated in administering tests? Yes_____ No_____

If yes, (a) for how long have you done so? Years_____ Months_____

(b) have you ever found it appropriate to give help to someone taking the test if they could not continue without help on a particular item? Yes_____ No_____ Don't know_____

* 11. Measuring reliability. That is--determining if a test will give similar scores when measuring similar performance. For example, a person taking equivalent versions of the same test should score about the same on both, if he has had no practice in between.

Have you been involved in measuring the reliability of tests? Yes___ No___

If yes, (a) how long have you been involved in measuring reliability?

Years_____ Months_____

(b) dc you compute coefficients of reliability?

Yes_____ No_____ Don't know_____

* 12. Evaluating validity. The test developer must determine whether the test is actually measuring what it is supposed to measure. Personnel who score high on the test should also perform very well on the task that test is supposed to measure, while those who score low should not be able to perform the task as well.

Have you helped to validate tests? Yes_____ No_____

If yes, (a) how long have you been doing so? Years_____ Months_____

(b) do you use content validity as opposed to predictive validity?

Yes_____ No_____ Don't know_____

13. Scoring. How are tests generally scored? Are norms set as standards using bell shaped curves, or are "go-no go" type standards used?

Norms_____ go-no go_____ . Other_____

_____

_____

_____

To what uses are the test scores put?

14. One might be using test results to compare student performance. Higher-scoring students might be considered for promotion for example, while those passing with a lower score might not be so considered.

Do you test results to compare students?

Yes_____ No_____

If yes, (a) how long have you used test scores for comparisons?

Years_____ Months_____

(b) if a student doesn't get a passing score the first time, is he tested again? Yes_____ No_____ Don't know_____

15. Another use might be using test results to evaluate course adequacy. Sometimes the results of tests are used to evaluate the success of a course. Portions of a test that many students fail to perform well on are seen as reflecting a deficiency in the corresponding portion of a course. Courses can then be improved, using test results as feedback.

Have you used test results to help improve courses? Yes_____ No_____

If yes, (a) how long have you been doing so? Years_____ Months_____

(b) when you do so, are test criteria based on task objectives, rather than on course content? Yes____ No____ Don't know_____

16. Another use might be using test scores to diagnose areas in which students needed improvement.

Do you use tests for diagnostic purposes? Yes_____ No_____

If yes, how long have you been doing this? Years_____ Months_____

17. Are there other aspects of test development and use that you are aware of but I did not mention? Yes_____ No_____

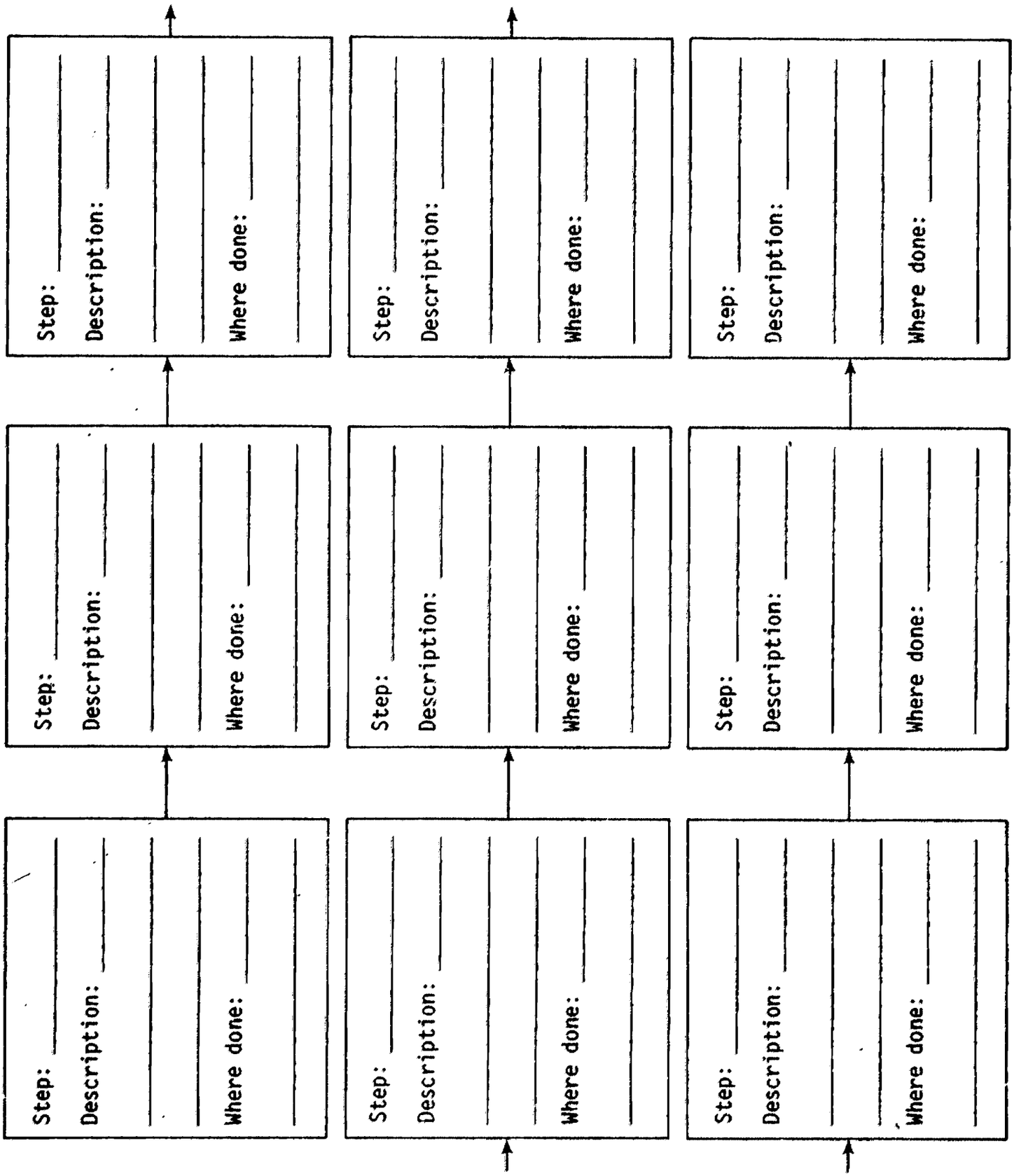
If yes, what are they?

_____

_____

_____

*18. Let's consider the overall test development and use process. Would you help me fill in the steps, as they actually happen at this post in developing and using tests? Since you may not participate in all steps yourself, we'd like to determine who does what step where.



Interviewer Statement: Now I would like to discuss some of the tasks that you're involved in.

19. What inputs do you have available in terms of documents, data, job aids, field manuals, etc.? REQUEST THESE

20. Which of these inputs do you actually use?

*21. [If answer to 20 is other than "all of them", interviewer asks #21]
Why do you use these and not the others?

22. What products do you prepare? REQUEST THESE

23. How are these outputs used?

24. What problems have you encountered?

25. How did you resolve these problems?

*26. **Is any special training available for testing personnel? Yes____ No____**

If yes, please briefly describe this training?

_____

_____

_____

27. What proportion of the tests you have participated in making or using are:

A. Paper-and-pencil knowledge tests? _____

B. Simulated performance tests? e.g., using mockups and drawings _____

C. "Hands on" performance tests? _____

D. Other? Specify: _____ _____

_____

What proportion of the tests you have participated in making or using are for:

A. Specific skill and knowledge requirements? _____

B. Specialty areas in a course? _____

C. End of block within a course? _____

D. Mid cycle within a course? _____

E. End of course? _____

*28. Are you familiar with any team performance situations that were evaluated by tests? Yes_____ No_____

*29. Would you briefly describe how tests were used to measure team performance?

_____

_____

_____

30. Have time pressures, or other constraints, prevented you from successfully carrying out some of the tasks involved in test construction and use?

Yes_____ No_____

If yes, describe how you were affected by a constraint.

_____

_____

_____

*31. **Can you describe any cases in which tests were developed which were not suitable, in your opinion, for the intended uses? Yes_____ No____**

Description: _____

_____

_____

If it is the interviewer's opinion that interviewee does not understand the distinction between Criterion-Referenced Testing and norm-referenced testing:

STOP HERE

Otherwise go on.

32. One of the main purposes of our work for the Army is to develop a manual on how to construct Criterion-Referenced as opposed to Norm-Referenced Tests. Who will be the primary users of a manual of this type on this post?

_____

_____

_____

*33. As you know, in recent years the Army has put increasing emphasis on using Criterion-Referenced Tests in appropriate testing situations. There is still much disagreement, though, about what a Criterion-Referenced Test really is. How is the term "Criterion-Referenced Test" used on this post?

_____

_____

_____

*34. How strongly do you feel about future use of Criterion-Referenced Testing in the Army? Should Criterion-Referenced Test development receive high or low priority in terms of Army assessment programs?

_____Strongly against--Criterion-Referenced Testing should receive bottom priority, or dropped entirely.

_____Against--Criterion-Referenced Testing should receive low priority.

_____Neutral--Criterion-Referenced Testing should receive average priority.

_____For--Criterion-Referenced Testing should receive high priority.

_____Strongly for--Criterion-Referenced Testing should receive top priority, Criterion-Referenced Tests should replace most or all norm-referenced tests.

*35. Do you think cost is a major factor in determining whether Criterion-Referenced Tests are developed and administered in the Army? That is--have you found that Criterion-Referenced Tests are more or less expensive to develop and administer than conventional, norm-referenced tests?

Less expensive_____ About the same_____ More expensive_____

*36. Could you describe a situation in which a Criterion-Referenced Test was found to be prohibitively expensive to develop?

_____

_____

_____

37. Do you think that there are any particular advantages or disadvantages to developing and using Criterion-Referenced tests in the Army (as opposed to norm-referenced measures)? Yes_____ No_____

What are some advantages or disadvantages?

_____

_____

_____

38. Are there any special problems you have encountered while developing or using Criterion-Referenced Tests, as opposed to problems normally encountered with norm-referenced tests? Yes_____ No_____

If yes, describe these special problems and how you overcome them:

_____

_____

_____

*39. How serious are these problems? That is, how much do they affect the overall accomplishment of testing objectives?

_____

_____

_____

40. Do you feel that Criterion-Referenced Testing is practical and useful in measuring job performance skills? Yes_____ No_____

Why?_____

_____

_____

*41. Are there other areas (such as knowledge tests and achievement tests) where this concept could be useful? Yes_____ No_____

Why?_____

_____

_____

42. What should we include to make the manual useful?

_____

_____

_____

APPENDIX B SUMMARY OF TYPES OF PERSONNEL INTERVIEWED AT ARMY INSTALLATION

Table B-1

FORT BENNING INTERVIEWEES

Classification Area	Directorate, Department or Division	Job Title of Interviewee	
U.S. Army Infantry School	Directorate of Educational Technology	Deputy Director	(S)**
	Faculty Development Division	Chief	(S)
		Senior Instructor	(DU)*
		Instructor	(DU)
		Instructor	(DU)
		Instructor	(DU)
		Student	(DU)
	Brigade & Battalion Operations Department (BBOD)		
	Operations & Training Techniques	Chairman	(S)
	Tactics Group	Test Officer	(S)
		Project Officer	(DU)
	Combat Support Group	Instructor	(DU)
		Instructor	(DU)
		Instructor	(DU)

**Supervisors of Test Development = (S)
*Test Developers or Users = (DU)

Table B-1 (continued)

Classification Area	Directorate, Department or Division	Job Title of Interviewee	
U.S. Army Infantry School (continued)	Directorate of Instruction	Chief	(S)
	Evaluation Division	Evaluation Staff	(DU)
	Curriculum Division	Director of Instruction	(S)
	Office of Directorate of Doctrine & Training		
	Task Analysis Division	Chief	(S)
	Training Management Team	Chief	(S)
	Office of Medical Staff & Operations		
	Instructional Division	Chief	(DU)
		Chairman, Resident Committee	(DU)
	Weapons Department Mortar Committee	Instructor	(DU)
TEC Program		Chief	(S)
MOS Testing Program		Chief	(S)

Table B-2

FORT BLISS INTERVIEWEES

Classification Area	Directorate, Department or Division	Job Title of Interviewee	
U.S. Army Air Defense School	High Altitude Missile Department	Training Specialist	(S)**
		Chief Project Officer for Curriculum	(S)
		Training Specialist	(DU)*
	Missile Electronic & Control Systems Department	Technical Publications Editor	(S)
		Instructor	(DU)
	Command & Staff Department	Chief, Command & Leadership Division	(S)
		Instructor	(DU)
		Department Staff	(DU)
	Army-wide Training Support Division	Educational Specialist	(DU)
		Educational Specialist	(DU)
		Assistant Chief of Course Development	(DU)
	Low Altitude Air Defense Department	Instructor	(DU)
		Instructor & Technical Writer	(DU)
		Department Staff	(DU)
	Ballistic Missile Defense Department	Training Specialist	(DU)
		Instructor	(DU)
	Deputy Commandant for Training & Education	Executive Officer	(S)
		Staff	(S)

**Supervisors of Test Development = (S)

*Test Developers or Users = (DU)

Table B-2 (continued)

Classification Area	Directorate, Department or Division	Job Title of Interviewee	
U.S. Army Air Defense School (continued)	Office of the Commandant	Education Advisor	(S)
TEC Program	Training Development Division	Chief of the Division	(S)
		Chief Project Officer for TEC Production	(S)
		Project Officer	(DU)
		Project Officer	(DU)
Training Center Program	Air Defense Artillery Training Brigade	Training Coordinator	(DU)
		Instructor	(DU)
		Evaluator	(DU)

Table B-3

FORT SILL INTERVIEWEES

Classification Area	Directorate, Department or Division	Job Title of Interviewee	
U.S. Army Field Artillery Training School	Tactic Combined Arms Department	Chief, Associate Arms Division	(S)**
		Senior Instructor	(DU)*
	Gunnery Department	Chief, Exam Branch	(S)
		Instructor/Grader	(DU)
	Office of the Commandant	Education Advisor	(S)
	Office of the Deputy Assistant Commandant for Training & Education	Educational Specialist	(S)
		Educational Specialist	(S)
	Materiel & Maintenance Department	Chief, Cannon Division	(S)
		Instructor	(DU)
	Target Acquisition Department	Supervisory Training Specialist	(S)
		Instructor	(DU)
	Command, Leadership and Training Department	Senior Instructor	(DU)
		Senior Instructor	(DU)
	Communications/Electronics Department	Training Instructor	(DU)
MOS Testing Program	Evaluation Brigade	Chief, MOS Analysis	(S)
Training Center Program	Advanced Individual Training Brigade	Officer in Charge	(S)
		Senior Instructor	(DU)
		Instructor in Charge of NCOs	(DU)

**Supervisors of Test Development = (S)
*Test Development or Users = (DU)

Table B-3 (continued)

Classification Area	Directorate, Department or Division	Job Title of Interviewee	
TEC Program	Army-Wide Training Support Department	Chief of Department	(S)
		Chief, TEC Branch	(S)
		Educational Specialist	(DU)

Table B-4

FORT KNOX INTERVIEWEES

Classification Area	Directorate, Department or Division	Job Title of Interviewee
U.S. Army Armor School	Directorate of Training	Chief, Task Analysis Division (S)**
		Test Director, MOS Evaluations (S)
	Leadership Department	Instructor, System and Procedures Branch (DU)*
	Army Wide Training Support	Chief, Development Division (S)
	Directorate of Instruction	Chief, Instructi . Technology Divisi.n (S)
		Instructor, Instruction Technology Division (DU)
		Educational Specialist, Evaluation Branch (S)
		Chief, Curriculum Branch (S)
	C and S Department	Chief, Cavalry Branch (DU/S)
		Senior Instructor, Small Unit Tactical Operations (DU)
	Automotive Department	Chief, Quality Control Branch (S)
	Weapons Department	Training Administrator (DU)
Training Center	Headquarters 1st AIT Brigade	S-3 1st AIT Brigade (S)

**Supervisors of Test Development = (S)
*Test Developers or Users = (DU)

Table B-5

FORT ORD INTERVIEWEES

Classification Area	Directorate, Department or Division	Job Title of Interviewee	
U.S. Army Training Center			
Directorate of Plans and Training	Quality Control Branch	Chief, Quality Control Branch	(S)**
		Training Evaluator, Quality Control Branch	(S)
	Basic Combat Training Testing	Project Test Officer, Quality Control Branch	(DU)*
		Instructor, Proficiency Test Branch	(DU)
Basic Combat Training	Training Command (Prov)	Operations and Training Officer	(S)
	Training Brigade	Battalion Commander	(S)
		Battalion Executive Officer	(S)
		Company Commander	(S)
		Company Commander	(S)
		Officer-in-Charge, First Aid Committee Group	(DU)
		Instructor, First Aid Committee Group	(DU)

**Supervisors of Test Development = (S)
*Test Developers or Users = (DU)

Table B-5 (continued)

Classification	Directorate, Department or Division	Job Title of Interviewee
Basic Combat Training (continued)		Noncommissioned Officer-in-Charge of Individual Tactical Training (DU)
		Senior Drill Instructor (DU)
		Drill Instructor (DU)
Advanced Individual Training	Field Wireman Division	Chief, Field Wireman Training Division (S)
		Instructor, Field Wireman Training Division (DU)
	Food Services Division	Supervisor, Food Services Division (S)
		Instructor, Food Services Division (DU)

APPENDIX C QUANTITATIVE DATA GATHERED DURING ARMY CRT SURVEY

Fort Benning, Georgia

Item No.	School Curriculum S* # Resp.	School Curriculum S* % Yes	School Curriculum DU** # Resp.	School Curriculum DU** % Yes	Training Center Curriculum S # Resp.	Training Center Curriculum S % Yes	Training Center Curriculum DU # Resp.	Training Center Curriculum DU % Yes	MOS Testing Branch S # Resp.	MOS Testing Branch S % Yes	MOS Testing Branch DU # Resp.	MOS Testing Branch DU % Yes	TEC (Training Extension Course) S # Resp.	TEC (Training Extension Course) S % Yes	TEC (Training Extension Course) DU # Resp.	TEC (Training Extension Course) DU % Yes
4. Involved in writing objectives?	3	100	12	88					1	100	1	100	1	100	1	100
4b. Objectives operationally, behaviorally written?	1	0	6	67											1	100
5. Participated in setting objectives?	3	67	11	64					1	100	1	100	1	100	1	100
6. Imposed practical constraints?	3	67	12	83					1	100	1	100	1	100	1	100
7. Helped determine priorities?	2	50	12	58					1	100	1	0	1	100	1	0
8. Did you write items?	3	33	12	100					1	100	1	100	1	100	1	100
8b. Item pool?	3	0	11	73					1	100					1	100
9. Involved in selecting final items?	2	0	11	45					1	100	1	0	1	100	1	0
9b. Use item analysis technique?	1	0	8	25					1	100			1	0		
10. Participated in test administration?	3	66	12	100					1	0	1	100	1	100	1	100
10b. Ever assisted someone taking test?	2	50	12	83							1	100				
11. Involved in measuring test reliability?	3	33	11	55					1	100	1	100	1	100	1	0
11b. Compute coefficients of reliability?	2	0	4	50					1	100			1	100		
12. Aid in validating tests?	3	0	11	76					1	100	1	100	1	100	1	0
12b. Use of content validity?	2	50	3	0									1	100		
13. Scoring: Norm or go-no-go?	5	60	12	33					1	0	1	0	1	100	1	100
14. Test results used to compare student performance?	3	66	11	91					1	100	1	100	1	100	1	0
14b. Retest?	2	100	7	28												
15. Feedback used to improve tests?	4	50	12	50					1	0	1	100	1	100	1	100
16. Tests used for diagnosis?	3	33	12	25					1	100	1	100	1	100	1	0
17. Aware of other aspects?	2	50	5	40					1	0					1	0
26. Training available for testing?	2	50	9	77					1	100	1	0	1	100	1	100
27. See following page for this item.																
28. Tests for team performance evaluation?	2	100	8	75					1	100	1	0	1	100	1	0
30. Constraints restrictive to test development?	3	67	10	80												
31. Any tests unsuitable for intended uses?	1	100	4	50												
34. See following page for this item.																
35. Are CRTs more expensive than NRTs?	4	75	3	67							1	0	1	100		
46. Criterion-referenced testing practical and useful?	1	100	9	100					1	100	1	100			1	100

* Supervisors of Test Development and/or Use.
** Developers and/or Users of Tests.

Fort Benning, Georgia

Item 27

Proportion of tests made or used:

		# Resp.	A. Paper & Pencil Tests %	B. Simulated Performance Tests %	C. "Hands-On" Tests %	D. Other %
School Curriculum	S*	2	12.5	12.5	75	0
	DU**	11	73	0	27	0
Training Center Curriculum	S					
	DU					
MOS Testing Branch	S	1	100	0	0	0
	DU	1	50	0	50	0
TEC (Training Extension Course)	S	1	10	20	70	0
	DU	1	50	50	0	0

Item 27, Part 2

Proportion of tests made or used for:

		# Resp.	A. Specific Skill & Knowledge Requirements %	B. Specialty Areas in a Course %	C. End of Block Within a Course %	D. Mid-Cycle Within a Course %	E. End of Course %
School Curriculum	S	1	20	20	20	20	20
	DU	9	11	14	34	11	30
Training Center Curriculum	S						
	DU						
MOS Testing Branch	S						
	DU	1	20	20	20	20	20
TEC (Training Extension Course)	S	1	0	0	25	0	75
	DU						

Item 34

Strength of opinion about future of CRT in Army

		# Resp.	Strongly Against %	Against %	Neutral %	For %	Strongly For %
School Curriculum	S	4	0	0	0	50	50
	DU	10			30	50	20
Training Center Curriculum	S						
	DU						
MOS Testing Branch	S	1	0	0	0	0	100
	DU	1	0	0	0	00	0
TEC (Training Extension Course)	S	1	0	0	0	0	100
	DU	1	0	0	0	0	100

* Supervisors of Test Development and/or Use.
** Developers and/or Users of Tests.

Fort Bliss, Texas

Item No.	School Curriculum S* # Resp.	School Curriculum S* % Yes	School Curriculum DU** # Resp.	School Curriculum DU** % Yes	Training Center Curriculum S # Resp.	Training Center Curriculum S % Yes	Training Center Curriculum DU # Resp.	Training Center Curriculum DU % Yes	MOS Testing Branch S # Resp.	MOS Testing Branch S % Yes	MOS Testing Branch DU # Resp.	MOS Testing Branch DU % Yes	TEC (Training Extension Cour.e) S # Resp.	TEC (Training Extension Cour.e) S % Yes	TEC (Training Extension Cour.e) DU # Resp.	TEC (Training Extension Cour.e) DU % Yes
4. Involved in writing objectives?	7	71	12	100			2	100					1	100	2	100
4b. Objectives operationally, behaviorally written?	3	67	11	82			2	100								
5. Participated in setting objectives?	6	83	12	100			2	50					1	100	2	1[illegible]
6. Imposed practical contraints?	7	57	11	91			2	50					1	100	2	100
7. Helped determine priorities?	7	71	11	91			2	50					1	100	2	100
8. Did you write items?	6	83	11	82			2	0					1	0	2	100
8b. Item pool?	5	80	10	80											2	100
9. Involved in selecting final items?	5	60	12	83			2	0					1	0	2	100
9b. Use item analysis technique?	6	66	12	92									1	100	2	0
10. Participated in test administration?	7	71	11	82			2	100					1	100	2	0
10b. Ever assisted someone taking test?	6	83	11	73			2	100					1	0	2	100
11. Involved in measuring test reliability?	7	57	12	42			2	50					1	100	2	0
11b. Compute coefficients of reliability?	6	33	7	0			1	0					1	100	2	0
12. Aid in validating tests?	7	43	12	17			2	0					1	100	2	100
12b. Use of content validity?	5	60	6	17									1	100	2	100
13. Scoring: Norm or go-no-go?	7	14	12	50			2	100					1	100	2	0
14. Test results used to compare student performance?	7	100	11	54			2	50					1	100		
14b. Retest?	4	50	5	80			1	100					1	0		
15. Feedback used to improve tests?	7	43	12	92			2	100					1	100	2	100
16. Tests used for diagnosis?	7	72	11	72			2	100					1	100	2	100
17. Aware of other aspects?	4	100	12	42			2	0							2	100
26. Training available for testing?	4	100	9	88			2	100							2	100
27. See following page for this item.																
28. Tests for team performance evaluation?	7	71	12	50			2	50							2	0
30. Constraints restrictive to test development?	7	100	10	100			2	100								
31. Any tests unsuitable for intended uses?	1	100	7	29			2	0								
34. See following page for this item.																
35. Are CRTs more expensive than NRTs?	1	100	9	76			1	100					1	100	2	100
40. Criterion-referenced testing practical and useful?	7	100	10	100			1	100					1	100	2	100

* Supervisors of Test Development and/or Use.
** Developers and/or Users of Tests.

Fort Bliss, Texas

Item 27

Proportion of tests made or used:

		# Resp.	A. Paper & Pencil Tests %	B. Simulated Performance Tests %	C. "Hands-On" Tests %	D. Other %
School Curriculum	S*	4	15	0	85	0
	DU**	12	53	7	36	4
Training Center Curriculum	S					
	DU	2	10	0	90	0
MOS Testing Branch	S					
	DU					
TEC (Training Extension Course)	S	1	100	0	0	0
	DU	2	100	0	0	0

Item 27, Part 2

Proportion of tests made or used for:

		# Resp.	A. Specific Skill & Knowledge Requirements %	B. Specialty Areas in a Course %	C. End of Block Within a Course %	D. Mid-Cycle Within a Course %	E. End of Course %
School Curriculum	S	3	0	11	67	3	19
	DU	10	21	18	30	4	27
Training Center Curriculum	S						
	DU	2	50	0	50	0	0
MOS Testing Branch	S						
	DU						
TEC (Training Extension Course)	S	1	0	0	90	0	10
	DU	2	0	0	90	0	10

Item 34

Strength of opinion about future of CRT in Army

		# Resp.	Strongly Against %	Against %	Neutral %	For %	Strongly For %
School Curriculum	S	7	0	0	14	86	0
	DU	10	0	10	10	20	60
Training Center Curriculum	S						
	DU	2	0	0	0	0	100
MOS Testing Branch	S						
	DU						
TEC (Training Extension Course)	S	1	0	0	0	0	100
	DU	2	0	0	0	0	100

* Supervisors of Test Development and/or Use.
** Developers and/or Users of Tests.

Fort Sill, Oklahoma

Item No.	School Curriculum S* # Resp.	School Curriculum S* % Yes	School Curriculum DU** # Resp.	School Curriculum DU** % Yes	Training Center Curriculum S # Resp.	Training Center Curriculum S % Yes	Training Center Curriculum DU # Resp.	Training Center Curriculum DU % Yes	MOS Testing Branch S # Resp.	MOS Testing Branch S % Yes	MOS Testing Branch DU # Resp.	MOS Testing Branch DU % Yes	TEC (Training Extension Course) S # Resp.	TEC (Training Extension Course) S % Yes	TEC (Training Extension Course) DU # Resp.	TEC (Training Extension Course) DU % Yes
4. Involved in writing objectives?	3	100	4	75												
4b. Objectives operationally, behaviorally written?	3	100	1	100												
5. Participated in setting objectives?	3	100	2	50												
6. Imposed practical constraints?	3	67	2	50					1	100						
7. Helped determine priorities?	3	100	2	50												
8. Did you write items?	3	100	2	100												
8b. Item pool?	2	100	1	100												
9. Involved in selecting final items?	3	100	2	50												
9b. Use item analysis technique?	3	67	2	50												
10. Participated in test administration?	3	100	4	100												
10b. Ever assisted someone taking test?	3	33	4	25												
11. Involved in measuring test reliability?	4	25	3	0												
11b. Compute coefficients of reliability?	1	0	1	0												
12. Aid in validating tests?	4	50	3	0												
12b. Use of content validity?	3	33														
13. Scoring: Norm or go-no-go?	5	20	8	13	1	100	2	50	1	100			2	100	1	0
14. Test results used to compare student performance?	5	60	8	50	1	0	2	0	1	100						
14b. Retest?	4	25	5	100	1	100	2	100								
15. Feedback used to improve tests?	4	100	8	75	1	0	2	0	1	0						
16. Tests used for diagnosis?	4	75	8	73	1	100	2	100	1	100						
17. Aware of other aspects?	3	0	7	14												
26. Training available for testing?	4	100	4	100	1	100	2	100					2	100	1	100
27. See following page for this item.																
28. Tests for team performance evaluation?	4	50	6	17	1	100	2	100	1	100			2	100	1	0
30. Constraints restrictive to test development?	4	50	8	26	1	0	2	0	1	100			2	0	1	0
31. Any tests unsuitable for intended uses?	3	33	3	0	1	0	2	0	1	0			2	0	1	0
34. See following page for this item.																
35. Are CRTs more expensive than NRTs?	3	100														
40. Criterion-referenced testing practical and useful?	2	100	3	100					1	100						

* Supervisors of Test Development and/or Use.
** Developers and/or Users of Tests.

Fort Sill, Oklahoma

Item 27

Proportion of tests made or used:

		# Resp.	A. Paper & Pencil Tests %	B. Simulated Performance Tests %	C. "Hands-On" Tests %	D. Other %
School Curriculum	S*	5	49	8	29	13
	DU**	8	59	0	41	0
Training Center Curriculum	S	1	50	0	50	0
	DU	2	50	0	50	0
MOS Testing Branch	S	1	75	15	10	0
	DU					
TEC (Training Extension Course)	S	2	50	50	0	0
	DU	2	100	0	0	0

Item 27, Part 2

Proportion of tests made or used for:

		# Resp.	A. Specific Skill & Knowledge Requirements %	B. Specialty Areas in a Course %	C. End of Block Within a Course %	D. Mid-Cycle Within a Course %	E. End of Course %
School Curriculum	S	5	8	28	60	0	4
	DU	8	8	26	61	2	3
Training Center Curriculum	S	1	100	0	0	0	0
	DU	2	100	0	0	0	0
MOS Testing Branch	S	1	100	0	0	0	0
	DU						
TEC (Training Extension Course)	S	2	100	0	0	0	0
	DU	1	100	0	0	0	0

Item 34

Strength of opinion about future of CRT in Army

		# Resp.	Strongly Against %	Against %	Neutral %	For %	Strongly For %
School Curriculum	S	3	0	0	25	0	75
	DU	1	0	0	0	0	100
Training Center Curriculum	S	1	0	0	0	0	100
	DU	2	0	0	0	0	100
MOS Testing Branch	S	1	0	0	0	100	0
	DU						
TEC (Training Extension Course)	S	2	0	0	0	0	100
	DU						

* Supervisors of Test Development and/or Use.
** Developers and/or Users of Tests.

Fort Knox, Kentucky

Item No.	School Curriculum S* # Resps.	School Curriculum S* % Yes	School Curriculum DD** # Resps.	School Curriculum DD** % Yes	Training Center Curriculum S # Resps.	Training Center Curriculum S % Yes	Training Center Curriculum DD # Resps.	Training Center Curriculum DD % Yes	MOS Testing Branch S # Resps.	MOS Testing Branch S % Yes	MOS Testing Branch DD # Resps.	MOS Testing Branch DD % Yes	TTC (Training Extension Course) S # Resps.	TTC (Training Extension Course) S % Yes	TTC (Training Extension Course) DD # Resps.	TTC (Training Extension Course) DD % Yes
4. Involved in writing objectives?	6	50	6	50	1	100			1	100						
4b. Objectives operationally, behaviorally written?	2	50	3	67	1	100			1	0						
5. Participated in setting objectives?	6	67	6	67	1	100			1	100						
6. Imposed practical constraints?	6	50	6	17	1	100			1	0						
7. Helped determine priorities?	6	83	6	50	1	100			1	100						
8. Did you write items?	6	17	6	67	1	0			1	100						
8b. Item pool?	3	33	4	50												
9. Involved in selecting final items?	5	80	6	33	1	0			1	100						
9b. Use item analysis technique?	3	67							1	0						
10. Participated in test administration?	6	67	6	83	1	100			1	100						
10b. Ever assisted someone taking test?	4	50	4	50	1	100			1	0						
11. Involved in measuring test reliability?	6	50	6	33	1	0			1	100						
11b. Compute coefficients of reliability?	3	100							1	100						
12. Aid in validating tests?	6	50	6	33	1	0			1	100						
12b. Use of content validity?	2	50	2	0					1	100						
13. Scoring? Norm or go-no-go?	6	33	6	33	1	100			1	0						
14. Test results used to compare student performance	6	83	6	100	1	0			1	100						
14b. Retest?	5	60	6	83					1	100						
15. Feedback used to improve tests?	6	83	6	50	1	100			1	100						
16. Tests used for diagnosis?	5	80	6	100	1	100			1	0						
17. Aware of other aspects?	4	25	6	0	1	100			1	0						
26. Training available for testing?	6	100	6	83	1	100			1	100						
27. See following page for this item.																
28. Tests for team performance evaluation?	6	33	6	33	1	0			1	0						
30. Constraints restrictive to test development?	6	67	6	67	1	100			1	0						
31. Any tests unsuitable for intended uses?	5	40	6	50	1	0			1	0						
34. See following page for this item.																
35. Are CRIs more expensive than NRIs?	7	71	6	67	1	100			1	100						
40. Criterion-referenced testing practical and useful?	6	100	5	100	1	100			1	100						

* Supervisors of Test Development and/or Use.
** Developers and/or Users of Tests.

Fort Knox, Kentucky

Item 27

Proportion of tests made or used:

		# Resp.	A. Paper & Pencil Tests %	B. Simulated Performance Tests %	C. "Hands-On" Tests %	D. Other %
School Curriculum	S*	6	15	19	31	35
	DU**	6	27	8	27	38
Training Center Curriculum	S	1	0	0	100	0
	DU					
MOS Testing Branch	S	1	80	0	20	0
	DU					
TEC (Training Extension Course)	S					
	DU					

Item 27, Part 2

Proportion of tests made or used for:

		# Resp.	A. Specific Skill & Knowledge Requirements %	B. Specialty Areas in a Course %	C. End of Block Within a Course %	D. Mid-Cycle Within a Course %	E. End of Course %
School Curriculum	S	3	47	13	13	13	13
	DU	4	43	0	0	0	57
Training Center Curriculum	S	1	100	0	0	0	0
	DU						
MOS Testing Branch	S						
	DU						
TEC (Training Extension Course)	S						
	DU						

Item 34

Strength of opinion about future of CRT in Army

		# Resp.	Strongly Against %	Against %	Neutral %	For %	Strongly For %
School Curriculum	S	7	0	0	0	14	86
	DU	6	0	0	0	33	67
Training Center Curriculum	S	1	0	0	0	0	100
	DU						
MOS Testing Branch	S	1	100	0	0	0	0
	DU						
TEC (Training Extension Course)	S						
	DU						

* Supervisors of Test Development and/or Use.
** Developers and/or Users of Tests.

Fort Ord, California

Item No.	School Curriculum S* # Resp.	School Curriculum S* % Yes	School Curriculum DU** # Resp.	School Curriculum DU** % Yes	Training Center Curriculum S # Resp.	Training Center Curriculum S % Yes	Training Center Curriculum DU # Resp.	Training Center Curriculum DU % Yes	MOS Testing Branch S # Resp.	MOS Testing Branch S % Yes	MOS Testing Branch DU # Resp.	MOS Testing Branch DU % Yes	TEC (Training Extension Course) S # Resp.	TEC (Training Extension Course) S % Yes	TEC (Training Extension Course) DU # Resp.	TEC (Training Extension Course) DU % Yes
4. Involved in writing objectives?					5	80	7	28								
4b. Objectives operationally, behaviorally written?					3	67	4	25								
5. Participated in setting objectives?					3	100	6	33								
6. Imposed practical constraints?					4	75	6	50								
7. Helped determine priorities?					4	75	6	33								
8. Did you write items?					3	67	5	40								
8b. Item pool?					3	67	4	25								
9. Involved in selecting final items?					4	50	6	33								
9b. Use item analysis technique?					4	0	5	0								
10. Participated in test administration?					11	91	11	100								
10b. Ever assisted someone taking test?					11	64	10	90								
11. Involved in measuring test reliability?					11	0	10	0								
11b. Compute coefficients of reliability?					4	0	6	0								
12. Aid in validating tests?					12	17	12	0								
12b. Use of content validity?					7	43	6	0								
13. Scoring: Norm or go-no-go?					11	100	10	80								
14. Test results used to compare student performance?					10	80	11	64								
14b. Retest?					5	100	6	83								
15. Feedback used to improve tests?					11	100	11	91								
16. Tests used for diagnosis?					11	82	11	82								
17. Aware of other aspects?					11	36	10	10								
26. Training available for testing?					10	0	9	33								
27. See following page for this item.																
28. Tests for team performance evaluation?					9	11	11	9								
30. Constraints restrictive to test development?					13	46	11	45								
31. Any tests unsuitable for intended uses?					8	50	5	80								
34. See following page for this item.																
35. Are CRTs more expensive than NRTs?					4	50	4	25								
40. Criterion-referenced testing practical and useful?					5	100	6	100								

* Supervisors of Test Development and/or Use.
** Developers and/or Users of Tests.

Fort Ord, California

Item 27

Proportion of tests made or used:

		# Resp.	A. Paper & Pencil Tests %	B. Simulated Performance Tests %	C. "Hands-On" Tests %	D. Other %
School Curriculum	S*					
	DU**					
Training Center Curriculum	S	10	0	0	100	0
	DU	11	0	0	100	0
MOS Testing Branch	S					
	DU					
TEC (Training Extension Course)	S					
	DU					

Item 27, Part 2

Proportion of tests made or used for:

		# Resp.	A. Specific Skill & Knowledge Requirements %	B. Specialty Areas in a Course %	C. End of Block Within a Course %	D. Mid-Cycle Within a Course %	E. End of Course %
School Curriculum	S*						
	DU**						
Training Center Curriculum	S	8	0	14	41	29	16
	DU	7	0	0	27	42	31
MOS Testing Branch	S						
	DU						
TEC (Training Extension Course)	S						
	DU						

Item 34

Strength of opinion about future of CRT in Army

		# Resp.	Strongly Against %	Against %	Neutral %	For %	Strongly For %
School Curriculum	S						
	DU						
Training Center Curriculum	S	8	0	0	12	0	88
	DU	7	0	0	13	29	58
MOS Testing Branch	S						
	DU						
TEC (Training Extension Course)	S						
	DU						

* Supervisors of Test Development and/or Use.
** Developers and/or Users of Tests.