AD A072448

ARI TECHNICAL REPORT
TR-79-A10

Principles of Work Sample Testing: III. Construction and Evaluation of Work Sample Tests

by

Robert M. Guion

BOWLING GREEN STATE UNIVERSITY
Bowling Green, Ohio 43403

LEVEL III

April 1979

Contract DAHC 19-77-C-0007

Prepared for

U.S. ARMY RESEARCH INSTITUTE
for the BEHAVIORAL and SOCIAL SCIENCES
5001 Eisenhower Avenue
Alexandria, Virginia 22333

Approved for public release; distribution unlimited.

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

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

	WILLIAM L. HAUSER
JOSEPH ZEIDNER	Colonel, US Army
Technical Director	Commander

NOTICES

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

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

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

C-2

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: 19 TR-79-A10 — 18 ARI

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle): 6 PRINCIPLES OF WORK SAMPLE TESTING. III. CONSTRUCTION AND EVALUATION OF WORK SAMPLE TESTS

5. TYPE OF REPORT & PERIOD COVERED: 9 Final rept. 15 Nov [redacted]76 - 15 Jun[redacted]78

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s): 10 Robert M./Guion

8. CONTRACT OR GRANT NUMBER(s): 15 DAHC19-77-C-0007

9. PERFORMING ORGANIZATION NAME AND ADDRESS: Bowling Green State University, Bowling Green, Ohio 43403

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 16 2Q161102B74F

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

12. REPORT DATE: 11 Apr[redacted]79

13. NUMBER OF PAGES: 89

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): -- 12 98p.

15. SECURITY CLASS. (of this report): Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

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

--

18. SUPPLEMENTARY NOTES

Monitored by G. Gary Boycan, Engagement Simulation Technical Area, Army Research Institute.

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

Measurement theory, psychometrics, work sample testing, validity, content-referenced testing, criterion-referenced testing, latent trait theory, generalizability theory

061 200 Gm

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

Work sample tests should be relevant to the job, objectively constructed and scored, reliable, and capable of being scored on a standardized content-referenced scale. Detailed steps in working from job analysis to establishing test specifications are presented for assuring job relevance. Methods are suggested for developing scales for scoring by a priori scaling, or by latent trait analysis, to provide a standard, content-referenced scale for scoring. Job samples should be evaluated primarily in terms of relevance and of generalizability. Seven principles of work sample testing are offered to

20. (continued)

researchers.

This report, the third of four, is written for psychologists and others interested in research testing.

PRINCIPLES OF WORK SAMPLE TESTING: III. CONSTRUCTION AND EVALUATION OF WORK SAMPLE TESTS

BRIEF

Desirable characteristics of work sample tests include standard scales for content-referenced scoring, adequate variance for statistical analysis, objectivity in structure and scoring, reliability of both measures and classifications, functional unity in measurement, and job relevance.

To assure job relevance, a rather deliberate process of domain definition is recommended. It begins with a thorough definition of the total job, a job content universe, defined in terms of component tasks, checklist-defined task elements, and clusters of similar component tasks called task categories. From the total universe, a sample (non-random) is selected defining the portion of the universe important for testing; this portion is the job content domain. All possible ways test tasks or items could be constructed and scored to sample performance in the job content domain define a test content universe, a universe of all possible admissible operations. Selecting from this universe, again with no attempt to be random or representative, is based on practical considerations of time, cost, and measurement feasibility and yields a portion of the possible universe which is the test content domain. From this domain, test specifications are prepared and, if the test is developed according to those specifications, either (a) it will be unarguably job relevant, or (b) the question of relevance will be directed toward the domain definitions rather than toward the test.

With test specifications established, a work sample test is developed by a panel of qualified experts who are knowledgeable about the job and who have also had some training in the preparation of test items. Items for work sample tests may include conventional written items testing job knowledge or practical items based on the total job, on some direct sample of the job domain, or an abstraction from the domain. Methods are presented for systematizing panel judgments of item relevance.

Scores on the test should, for a work sample, be interpretable in terms of the content rather than only with reference to norm groups. That is, work sample tests should be content-referenced rather than norm-referenced. Some alternative methods of scoring are presented An emphasis is placed on test development and scoring using latent trait theory, which provides a standard content-referenced interpretation that can be independent of the distribution of a given sample

of people and of the particular set of items chosen for testing a given individual. An alternative system involving a priori scaling is also discussed. The practice of reducing all scores to a pass-fail dichotomy is discussed (procedures for doing so are briefly presented) and discouraged. For work sample tests, level of mastery seems more important from a measurement perspective than classification as a master; degree and direction of misclassification is more important than the mere fact of misclassification.

Generalizability theory is briefly discussed in this report, although a more detailed discussion of it is reserved for the fourth report of the series.

The evaluation of a work sample test should emphasize job relevance and generalizability. Other evaluations to be considered include reliability (with perhaps greater emphasis on the reliability of the scores as measures than on the reliability of classifications), accuracy of measurement, information curves, construct validity (with an emphasis on examining alternative explanations for scores), and in some cases evidence of predictive or other criterion-related correlations.

The paper concludes with seven basic principles of work sample testing: (1) choices of job content domains need justification, (2) test content domains should be as congruent as possible, (3) scoring procedures should strive toward fundamental measurement, emphasizing transitivity within a reasonably homogeneous domain, (4) scores should permit assessment of levels of proficiency rather than mere dichotomies, (5) opportunities for irrelevant influences on individual scores should be minimized, (6) scoring of work sample tests designed for use in large, multi-location organizations, should be standardized on a content-referenced scale applicable to the organization as a whole, and (7) scores on a work sample test given in a setting of institutional control and standardization should generalize to a variety of field settings.

TABLE OF CONTENTS

79 08 06 036

TABLE OF CONTENTS

LIST OF FIGURES

INTRODUCTION

At least since World War I (Yerkes, 1921), paper-and-pencil testing has been accepted as the prototype for psychometric theory. The massive testing programs initiated and carried out by the Army during that war were expanded in military services and in employment settings during the years between wars. By World War II, psychological testing for classification purposes, primarily using paper-and-pencil tests, was widespread. The Aviation Psychological Program (Guilford, 1947) used a few apparatus tests, such as measures of coordination and dexterity, but these were oddities in the general classification test batteries. Since World War II, testing practice has continued to be dominated by paper-and-pencil tests, mainly measures of knowledge or of cognitive variables.

Concomitantly with the growth of the use of the paper-and-pencil tests was growth in the development of the psychometric theory. Beginning with Thurstone's famous monograph on the primary mental abilities (Thurstone, 1938), through the factor analytic work in the Aviation Psychology Program, the publication of the monumental textbook by Lord & Novick (1968), and a substantial literature since, the theory of mental tests scores has moved from a few brief equations into an astonishingly complex mathematical structure.

No corresponding effort has been expended in the measurement of performance variables. Performance testing also dates back to before World War I, but in the Army, these tests were limited pretty much to testing for intelligence, not proficiency at designated tasks. The Stenquist Test of Mechanical Proficiency (requiring the assembly of common objects) was used as a group test of intelligence for "illiterates and foreigners." It was replaced by the beta examination.

"The chief objection to it was its low value as a measure of intelligence. Even with unselected literate men it correlated with examination a only to the extent of 0.45 to 0.55" (Yerkes, 1921, p. 321). If such tests had been evaluated as measures of performance variables rather than as measures of cognitive variables, and if the level of research and theory given to performance measurement even partially matched the work done on cognitive measurement, the potential value of measurement by work samples might have been realized long ago. Work samples yield more nearly fundamental measurement with less reliance on norms and with potentially less likelihood of contaminating sources of variance. In contemporary society, a major problem with paper-and-pencil testing is the charge of bias. Since bias seems largely due to irrelevant variance, performance variables measured directly by performance techniques may either show less evidence of discriminatory impact against women and minorities or be taken as evidence that observed group differences are real.

EXAMPLES OF WORK SAMPLE TESTING

A work sample test is defined here as any standardized and scorable procedure in which people are asked to answer questions, solve problems, produce or modify objects, or otherwise demonstrate knowledge and competence in tasks drawn from a job content domain. This is a broad definition. It is broad enough to include as a work sample test a systematic set of probationary assignments on which performance is systematically evaluated or "scored." It can also include, as a different extreme, a paper-and-pencil test of the knowledge identified as part of a job content domain. It can include literal job assignments, or simulations that faithfully reproduce aspects of such assignments, or abstractions from job assignments that appear to be artificial but reproduce essential or crucial aspects of an assignment.

It includes tests developed as criterion measures, training devices, predictors, or standards for certification. However, it does not include tests that involve varieties of common tasks, even manipulative tasks, unless performance on those tasks can be shown to be part of a specified job content domain. By this definition, a typical test of typing speed and accuracy, for example, does not become a work sample test until the job content domain it samples is specified.

The variety of possible kinds of work sample proficiency tests is limited only by the variety of jobs and the imaginations of the test developers. Asher and Sciarrino (1974) cited over eighty published accounts of work sample tests -- with virtually no overlapping examinations.

A six-year study of bias conducted by the Educational Testing Service in cooperation with the United States Civil Service Commission (Campbell, Crooks, Mahoney, & Rock, 1973) used work sample criteria for two of the three occupations studied. An attempt to develop a work sample for medical technicians was abandoned as unsuccessful. However, three kinds of work samples were successfully developed for Cartographic Technicians: one involving the compilation of contour lines, another extracting drainage system and cultural detail from vertical aerial photographs, and the third requiring a geometric restitution of information from oblique aerial photographs. An In-basket exercise was developed for Inventory Managers; it simulated decisions and communications concerning inventory following the Military Standard Requisitioning and Issue Procedure (MILSTRIP). A hypothetical new agency was "created" for the exercise.

Some work samples involve a great deal of ingenuity in their construction. Rubinsky and Smith (1973) simulated a bench grinder's job for an experiment in safety training. The simulation was so organized that an "accident" was signaled by turning on water jets. If the operator was standing in the correct place, he was not sprayed ("injured"); otherwise, he got wet.

In what ranks high as an understatement, Root, Epstein, Steinheiser, Hayes, Wood, Sulzen, Burgess, Mirabella, Erwin, and Johnson (1976) described the background for REALTRAIN in these terms, "A combat environment, which involves the violent interaction of two mobile opposing forces who are out to destroy one another, is difficult to simulate" (p. 1). The REALTRAIN exercise, in its various stages of development, seems to have had applied a substantial degree of ingenuity in simulating the critical component of combat experience, the knowledge of whether one has killed or been killed. In the REALTRAIN exercise, an initial procedure used telescopic sights for identifying a number on the helmet of an opposing soldier. Upon reading the other soldier's number through the telescopic sight, the soldier fired a blank round and reported the number to a controller who, by radio, identified the casualty to a controller with the opposing force. Knowledge of results with this technique was accomplished in five to ten seconds (Schriver, Mathers, Griffin, Jones, Word, Root, & Hayes, 1975). In describing the REALTRAIN exercise, Uhlaner, Drucker, and Camm (1977) reported an adaptation using lasers to simulate the direct fire characteristics of a number of weapons. Firing a blank round keys the firing of the laser; the method reduces the discrepancy between accurately sighting a target and accurately hitting it.

An abstract work sample reported by Grant and Bray (1970) was developed originally as a training device called the Learning Assessment Program (LAP). The LAP abstracts skilled activities from seven levels of telephone installation craft work. The first four refer to fairly conventional kinds of telephone installation, while the last three are abstracted from higher level jobs. A person tested under the Learning Assessment Program could have up to three weeks to learn and demonstrate the seven levels of proficiency; scores were highest level completed, highest level passed, and time taken to complete the program.

In some jobs a simulation exercise can be required and organized so that the thought processes and judgments can be tested using multiple-choice test items. An example is the examination procedure of the National Council of Architectural Registration Boards (NCARB, 1976). In its 1975 examination, the task was to design a performing arts center for Scottsdale, Arizona. The four-part, two-day examination included detailed information about the community, the needs, legal requirements, and other matters. A variety of environmental, programming, design, and construction questions were asked. The examination materials, including actual items, was subsequently published (NCARB, 1976) along with the announcement that the 1976 examination would center on designing a facility for a prison infirmary and health center. Approximately eight pages of advance information about the examination were presented along with a substantial bibliography of useful background books and articles.

An architect's work is essentially information processing; the quality of his final product is a matter of taste, but the necessary thought processes leading to that product are objectively known; a paper-and-pencil test can therefore be a satisfactory work sample. In some jobs, however, excellence in the work process is rather irrelevant if the outcome is poor. Also, some jobs are so designed that an individual cannot do them alone; the measurement of proficiency must use a work group rather than an individual as the unit of analysis.

An example of both problems is tank warfare. An excellent knowledge of battle theory or weapon nomenclature is of little value if the tank crew cannot hit a target and makes itself vulnerable to being hit; moreover, hitting targets and staying protected is not a one-man job. A proficiency test of tank crew gunnery has been developed and reported by Wheaton, Fingerman, and Boycan (1978). It consists of a set of simulated test engagements involving different types of targets, different required behaviors of individual crew members, and some practical constraints on the use of main gun ammunition.

Work sample tests, with an emphasis on job knowledge, are used in many state or trade licensing or certification programs. They vary not only in content but in quality. Shimberg, Esser, and Kruger (1973) have reviewed licensing practices and policies in different states and in different occupations, and they present a dismaying picture. Most such tests are written examinations; Shimberg, et al. identify problems with written examinations under four headings: lack of planning, over-reliance on unreliably scored essay tests, poor quality where multiple-choice questions are devised, and lack of item analysis. Even the performance portions of such tests are frequently inadequate because of failure to sample crucial skills adequately, the failure to standardize procedures, and the lack of reliable or appropriate scoring procedures for evaluating performance.

DESIRABLE CHARACTERISTICS OF WORK SAMPLES

A work sample test is an operational definition of proficiency in some aspect of performance of the work sampled. The requirements for evaluating a test simply as a satisfactory operational definition of proficiency level, as outlined in the second report in this series, are always important in work sample testing. If these have been satisfied, it seems unnecessary in most cases to consider still further requirements.

In many certification testing situations, however, some further problems intrude themselves. These situations include state or trade association certification examinations, or qualifying examinations for promotions, where the examinations are to be administered repeatedly or at various times and places in a multi-location organization. The Army skill qualification testing program is an example of the latter. The organization is by no means small, the jobs occur in many locations throughout the world, and the necessity to certify qualifications is a continual one. In these circumstances, there are some additional questions to be considered in the construction and evaluation of work sample tests.

STANDARD CONTENT-REFERENCED SCORES

Scores on a work sample should be interpretable in terms of test content and its relationship and the interpretation should be standard across examiners, locations, times, and conditions. There are several ways to standardize interpretations of scores. The simplest, although it is the most difficult to defend, is to interpret each score as either above or below an arbitrary cutoff point classifying examinees either as masters or as non-masters. Much of the literature on criterion-referenced testing in education seems bent on such a loss of information, and loss of information is the best description of most dichotomous scoring. It is much more useful to refer to the level or degree of mastery, a polychotomous scoring system.

When we think of standard scores, we typically think of the *z*-score scale, with its mean of zero and a standard deviation of 1, or of a linear or nonlinear transformation of it. Many linear transformations have been used; nonlinear transformations are ordinarily chosen in an effort to normalize obtained distributions. Examples include the familiar stanine or McCall's _T_-score with its mean of 50 and standard deviation of 10. These are the standard norm-referenced interpretations of scores, and they can be quite useful for many purposes. They are, for example, extremely useful for interpreting scores on aptitude tests or on measures of personality or attitude variables.

They are not, however, desirable interpretations of scores for work sample performance. A _T_-score between 40 and 60, for example, means only that the examinee performs about like nearly everyone else, at least in the normative sample, but it tells nothing about his level of mastery. Some form of content-referenced scale is usually necessary to provide adequate meaning for a work sample test. At least three kinds of content-referenced scales can be devised.

1. Occasionally, a group of expert judges, considering the examination in detail, will arive at a system for establishing an arbitrary cutting point or standard above which mastery may be claimed. Where such a standard is established, scores can be interpreted in terms of linear distances from that standard point. This can be a useful scale, but its value depends on how widely the standard is accepted.

2. A priori scaling can provide a basic reference scale. If a subset of test components or items form a scale, then total scores can be interpreted with reference to that scale of selected items.

3. If latent trait analysis is used, the test can be scored on the basis of maximum likelihood estimates or other estimates along a "sample-free" scale of underlying latent ability.

There is a special advantage associated with the last two examples.

If a work sample can be scored on an absolute scale, whether this scale be calibrated by latent trait analysis or by older scaling techniques, it will solve a long-standing problem in criterion-related validation. If the criterion is measured on such a scale, then a predictor variable that has been used to select people can be revalidated despite sample homogeneity. As Peterson and Wallace (1966) pointed out, the use of a valid predictor often results in such restricted variance that evidence of validity can no longer be obtained by computing validity coefficients. If the criterion measure is a work sample, however, and can be predicted with absolute predictions, then follow-up validation is possible by computing the variance of the errors in prediction and comparing it to the overall criterion variance in the original validation study.

VARIANCE

There has been a substantial controversy over the importance of variance in work sample testing (Popham & Husek, 1969; Millman & Popham, 1974; Woodson, 1974). The issue seems to center on the observation that an educational achievement test showing that everyone in a class has mastered the curriculum objectives by the end of the term is not necessarily a bad test; the lack of variance may simply mean excellent teaching. It may also mean poor measurement. The effect of training seems generally to be an increase in individual differences, not the elimination of them. Absence of variance in <u>scores</u> ought not be confused with absence of variance in the underlying variable being measured by them.

Quite apart from theoretical issues, there are practical reasons for seeking variability in test scores. One is that levels of mastery cannot be identified without individual differences in scores. Another is that one can never be sure that low variance does not simply indicate an excessively easy test. Of course, it may be useful to minimize variance within groups while maximizing variance between groups. As Millman and Popham (1974) point out, this is a question of validity, and of criterion-related validity at that.

However, if this is one's essential purpose in testing, then the objective of test construction should be a bimodal distribution of scores. In the ideal case, the discrimination values of the items is said to be about .5 (if half of the total sample is classified as masters and the other half as non-masters), and the average item correlation is 1.0. If this were the case, of course, everyone would achieve either a perfect score or a zero score, and a single item would have done as well as the total test. Since the ideal is never achieved, different items represent replications, and the more realistic goal is to obtain as little overlap as possible between the distributions of masters and non-masters.

This seems to be, however, a shortsighted objective in work sample test construction. For most purposes, the distribution of scores should probably be somewhere between a rectangular and a normal distribution. Such a distribution contributes to versatility in test use so that the same investment provides a test for identifying different levels of mastery, for validating aptitude tests, and for diagnosing organizational ills as well; moreover, its use can continue even if performance standards change.

PSYCHOMETRIC CONSIDERATIONS

Work sample tests tend to be reasonably objective. However, their objectivity, as defined in the preceding report of this series, can be spoiled by distortions in response, characteristics due to restrictive or ambiguous format, poorly motivating testing conditions, or unreliable scoring. Where observers are used, scoring procedures should be defined so completely that high levels of interscorer agreement, or conspect reliability, can be achieved.

It is useful to distinguish the reliability of measurement from the reliability of classification. If the test is to be used for classifying people into those who are certified and those who are not, or into multiple categories, then the reliability of the classifications achieved is an issue of importance. It is, however, an issue independent of the reliability of measurement as such. More will be said on this in the next major section; at the present time, the emphasis is on the reliability of measurement.

That is, the emphasis is on attempts to assure, at the very least, a minimal effect of random errors of measurement. This is the essence of the classical concept of reliability: the extent to which a set of measurements is free from random error variance.

The classical definition of reliability is minimal. Where one wants to generalize the inferences from test scores to "real-world" inferences, assessment of random error of variance yields a necessary but insufficient evaluation. A more important and more general concept is freedom from irrelevant sources of variance, which is the classical concept of validity. There are systematic errors of measurement as well as random errors; in the construction and evaluation of work sample tests, one should be particularly alert to sources of systematic error.

Moreover, one should be alert not only to sources of error variance, but also to sources of error in individual scores. Since the score on a work sample test is likely to be interpreted in terms of a standard or underlying scale, the test may be useful even in the isolated case where the proficiency of only one person is measured. A satisfactory measuring instrument could be used only one time, for only one person, yielding a score interpretable in terms of test content. This can only happen, however, if the tester can evaluate the degree to which that one use of the test is free from irrelevant kinds of error, even in the absence of a group for which to compute variance.

PREVIEW

In the sections that follow, procedures will be proposed for determining domains, establishing test content, scaling test components, and studying the generalizability of scores. In each section, the different implications for tests of job knowledge, literal work samples, or abstractions of literal work assignments will be discussed. All of this will be offered for the "ideal case" -- not for routine testing. In a given testing situation, some of the recommended procedures will be superfluous. In some situations, certain of these procedures will not be feasible.

Despite the differences between generally idealistic prescriptions and the realistic requirements of specific situations, there is a value in being unabashedly idealistic: it provides a conceptual standard against which one can assess the importance of deviations from the ideal in real cases. Without such a conceptual standard, the gradual progress toward improvement -- which is an attainable ideal -- is unlikely to occur.

DOMAIN DEFINITIONS

Ordinarily, the terms content universe and content domain are used interchangeably. In these reports, however, they have been distinguished; a domain is defined as a not-usually-random sample of a universe. The distinction is made because it offers a useful guide for thinking about judgment processes. It is not intended as a necessary and formal requirement for effective content sampling; nevertheless, the procedures implied by the distinction seem useful, particularly in job analysis.

Although job content domains and test content domains have some different elements in their definitions, both definitions begin with job analysis. The analysis of a job into its component functions can take many forms; the approach outlined here, including the use of the results of the analysis, is suggested primarily as a procedure assuring job relevance of the final test specifications. It begins with a global definition of a job content universe and ends with the specifications for test construction.

Three terms require definition:

1. A component task is a preliminary statement, in rather broad terms, of a major activity, task, or responsibility of the job. It may be an appropriately formalized sentence such as "__(Takes action)__ in __(setting)__ when __(action cue)__ occurs, using __(tools, knowledge, or skill)__."

2. A task element is a simple statement describing a detail of the component task; it may describe a movement, a judgment, a source of information, or some other aspect of the broader task. Task element statements may be arranged in the form of a checklist; the same task element may be part of more than one component task.

3. A task category is an empirically-identified group of component tasks sharing a common general pattern of task elements.

The precise nature of the formal sentence may vary according to the nature of the job. For example, Wheaton, et al. (1978), in their study of Tank Gunnery Tests, used a formula somewhat like this: "Given ________________ equipment and _______________ target under ________ conditions, a tank crew in ______________ position will open fire within _____________ seconds of the alert element of the command, and neutralize the target within _________ seconds using no more than ____________ rounds." Wheaton, et al. described such statements as "job objectives." Their terminology grows out of formal training for tank crew members and follows the language used by educational measurement specialists in writing about criterion-referenced testing. It is commonly asserted in that literature that the unit of analysis in content-referenced measurement is the instructional objective.

The broader term, task component, has been chosen here because content-referenced tests are used for many purposes other than assessing the outcome of instruction. Whether the job incumbent brings the ability to carry out a particular task with him, learns it in formal training, or picks it up on the job is of relatively little consequence in defining the nature of the task. To be sure, in nearly all statements of component tasks are statements of things the incumbent is expected to be able to do; in that sense they are objectives. The present discussion, however, will attempt to evade the overtones of instructional objectives by referring to component tasks.

One component task for an electrical appliance repair person might be "Repairs television set in customer's home at customer's request, using tools in portable kit and knowledge of circuits in that model and general electronic knowledge". A task element within that component task might include checking vacuum tubes for conformity to specifications. A given task element, such as checking tubes, may show up in other component tasks; it may, for example, be necessary to check tubes in servicing certain electronic air filters. Another task element, "disconnect line cord," would apply to virtually every repair job that he may undertake.

A subjective analog of the task category is the Army's "duty module concept" (Duffy, 1976). A closer example of what is implied by the term task category is provided by Wheaton, et al. (1978) in their cluster analysis of job objectives or, to use the terminology here, preferred task components.

It should be understood that the cluster analysis identifying a task category is not a cluster analysis of checklist items. In the language of Tryon and Bailey (1970), the analysis of checklist statements is a V-analysis, a clustering of variables. The identification of task categories, on the other hand, uses what those authors called an O-analysis, a clustering of objects according to common profiles. The objects, in this case, are component tasks; component tasks with the same profiles of task elements are, for practical purposes in testing, essentially similar tasks, and they may as well be assembled under a common heading.

DEFINITION OF JOB CONTENT UNIVERSE

Job analysis may begin by interviewing job incumbents or their supervisors, alone or in a group, to develop a set of component tasks. The procedure can be expedited by a list of action verbs that can begin the stylized statements of categories. Examples of appropriate action verbs include "decides," "repairs," "inspects," "assembles," "lubricates," and so forth. The use of key action verbs is more effective if there is a standard set of such verbs, as there is in the Air Force (Foley, 1977). Such a list, however, is not necessary, at least in a group interview, because the crucial list for a given job can be developed on an ad hoc basis. Once the verb for the formula sentence has been selected, the other blanks can be filled in relatively easily, and the first sentence is the hardest.

A job with a large number of component tasks may require a second round of interviews to verify the information obtained in the first. The process of verification is probably less concerned with checking the accuracy of the earlier statements than with stimulating the development of additional ones or of combining those for which the differences are obviously trivial.

When the component tasks have been completely identified, task element statements may be written and assembled into job description checklists. As a general rule, a task-oriented checklist provides a more direct set of specifications for work sample test construction than does a worker-oriented checklist. The latter is very useful in identifying predictor variables, training content, or the essential processes comprising a component task, but it seems less useful in defining the principal activities of the work sample.

Each component task may require several kinds of task element statements. Small panels of job incumbents or supervisors should probably do the initial writing. It may be better if they work as a committee rather than individually if the group activity will stimulate thinking. They should write task element statements under a number of different headings. One heading might be sources of information or material used in carrying out the task. Another heading might involve work processes: physical activities, perceptual judgments, or cognitive tasks expressed as specific decisions or judgments to be made. Specific actions or accomplishments can be another category, and still another might include prerequisite knowledge utilized in deciding on courses of action or carrying out required actions.

Despite repetition, elements should be listed under each component task. Each statement should be presented with a response scale for the evaluation of the importance of the elements, their duration or frequency of occurrence, percentage of total work time or level of skill they require, or any other scale the panel of experts deems appropriate. The experience of the present writer is that it makes very little difference what scale is used to describe the task elements since the various scales correlate very highly. The important consideration is that members of the panel themselves, and the workers they represent, will feel comfortable in using the scale chosen.

After a pilot study to assure clarity and completeness, the job description checklist should be administered to a large sample of incumbents or supervisors. The sample completing the checklist should be representative of potential diversities within the organization. If the job appears, for example, in regionally scattered installations or in installations varying markedly in size, then incumbents from each of the regions or each installation size should be surveyed. The survey can usually be done by mail, although interviews may be helpful if the checklist is complicated.

Data from the survey should be analyzed to identify the most common elements making up individual component tasks across sample characteristics such as region or installation size. Some task elements may be an actual part of the component task for only some respondents. Elements of a component task that are idiosyncratic to individuals or subgroups do not really define the task.

One might do a cluster analysis of all task elements and determine cluster scores for each component task by averaging across respondents. Each component task will then have a profile consisting of the same number of points as clusters in the checklist. If there has been sufficient uniformity of checklist items from one component task to another, these profiles can be compared in developing task categories. Clusters of component tasks with similar profiles identify task categories. The job content universe consists of the set of task categories, each defined by its own characteristic profile of task elements.

In developing the Tank Gunnery Test, Wheaton, et al. (1978) identified a list of 266 job objectives (component tasks) and 114 behavioral elements (task elements) of those "objectives." Assigning a dichotomous classification of either one or zero for each cell of the matrix (based on a restrictive assumption of a first-round hit), indicating whether the behavioral element was in fact involved in the job objective (component task), they cluster analyzed that matrix. The results of the analysis identified 16 relatively homogeneous clusters in terms of the behavioral elements that adequately defined the domain of interest if not the total job universe.

DEFINITION OF JOB CONTENT DOMAIN

The complete list of component tasks, each of which is defined in terms of its own list of task elements, offers a thorough definition of a job content universe. The definition of the job content universe is redundant and can be simplified without serious information loss to whatever degree component tasks may be clustered into task categories. Further redundancy exists because certain task elements are repeated far more frequently than others across the various categories. A random sample from the job content universe would, therefore, produce an unnecessarily redundant examination. Moreover, some parts of that universe may be trivial for the purposes of testing.

The definition of a job content domain is a matter of sampling from the job content universe, but it is by no means a matter of random or representative sampling. The sampling strategy should fit predetermined objectives. In the gunnery tests, for example, Wheaton, et al. developed a sampling strategy based on an "index of generalizability." Generalizability in this case is a function of the number of behavioral elements (task elements) in one component task that are also included in one or more other component tasks. They had decided that there would be proportional representation from each of the task categories according to the number of component tasks they contained. If proportionality suggested that one single component task would be chosen from a given category, it would be the one with the largest index of generalizability. If proportionality required two component tasks, then those with the largest indices of generalizability were chosen, and so on.

Several judgments such as these are made in defining a job content domain, and they should be made by a panel of experts consisting of job incumbents, their supervisors, or both. (Systems experts, industrial engineers, or others may be useful for some panels.) For a test designed to certify competence, such a panel needs to determine which kinds of task elements are the essential or critical elements in the various task categories. Decisions made at this point will greatly influence subsequent decisions whether to develop a test measuring hands-on performance, a written test of job-related decisions and judgments, or a job knowledge test. These decisions will not necessarily come easily; the test developer may expect sometimes acrimonious debate between those who believe that nothing matters but the real results of work and those who believe that good results are attributable only to dumb luck in the absence of well-informed judgment. Somehow, however, the different opinions must be reconciled and a consensus reached about the criticality of the task elements. What seems to be important in making these judgments is that strategies for selecting the sample of component tasks or task elements from the total universe be developed clearly, be accepted by the panel of experts who must make subsequent judgments, and be reasonable in the light of the objectives of work sample testing.

DEFINITION OF TEST CONTENT UNIVERSE

In his chapter on test validation, Cronbach (1971) referred to the "universe of admissible operations." Later, in the monograph on generalizability theory, he and his colleagues referred to the "universe of admissible observations" as the basis for domain definition (Cronbach, et al., 1972). The terms are doubtless interchangeable, but a possible difference of emphasis that makes a bit of word play instructive for the definition of work sample universes and domains.

<u>Operation</u> seems to imply doing something; admissible operations might be the acceptable things test builders do in providing the stimulus material or that test takers do as responses to that material -- the stimulus-response content of the test. <u>Observation</u> seems in addition to imply noting and evaluating what the test taker does -- the scoring content of the test. For work samples, at least, it seems the more inclusive term.

The Universe of Admissible Observations. It is very difficult to imagine a work sample performance test that does not involve measurement by direct observation. The measurement of such a test is the result of an evaluation either of the procedures used in carrying out the requisite tasks or of the evaluation of the results of task performance. In either case measurement depends on some form of observation or an observational aid. It is unlikely that credence could be given to descriptive accounts by job incumbents about the work processes, not because of possible faking so much as because jobs tend to become automatic. People who know how to do a job well simply do not know for sure what they do. If it is a product of the work that is to be evaluated, some form of inspection is needed. Inspecting is careful observing. It may be aided by various kinds of physical instrumentation ranging from nothing more complicated than a ruler to massive equipment for testing stresses or breaking points, but the final responsibility for judgment in the inspection of a product rests with the observer; ultimately, therefore, the measurement of job proficiency through the evaluation of the products of work is a form of direct observation.

A definition of a test content universe must specify both the stimulus materials (the assignments or questions) and the general form of probable or permissible responses (performance or answers). If it is a work sample test, the stimulus-response content must be drawn from the job content domain. The scoring content, however, does not exist in the job content domain, at least not ordinarily. It must be an added factor, and the emphasis on watching and scoring implied by speaking of observations makes it useful to define a test content universe as a universe of admissible observations.

In their generalizability monograph, Cronbach et al. (1972) point out that any observation can be described in the context of certain conditions. These may include such considerations as the nature of the task, the environmental setting in which it is performed, the time of day, or the level of external control over performance among many other possible considerations. Each of these is a *facet* of the observation, and there can be limits to the range of conditions within each facet of interest. In Army testing for skill qualifications, for example, environmental hostility can range from nearly none at all to combat conditions. One may decide on either technical or social bases that the *admissible* or acceptable set of environmental conditions for purposes of testing people may be limited to simulations of certain features of the combat condition.

The task components and some other facets of the universe of admissible observations are specified in the job content domain. These are not enough for testing. Other facets or conditions must be added to produce a generalizable set of scores. This is why a test content universe or domain cannot be equated with a job content domain. The added components must be added with careful, informed, and systematic judgment by qualified experts, and attempts should be made to consider, even if ultimately to reject, any potential element in the universe of admissible observations.

Item Forms. A useful guide to the development of a universe definition is the concept of the *item form* (Hively, Patterson, & Page, 1968). The research reported by Hively et al. described item forms for a test of basic arithmetic skills -- clearly a finite universe considerably smaller than the universe for ~~most~~ work-sample testing. However, the stylized sentence suggested for the identification of component tasks can be used to develop item forms for work samples. Additional elements are needed, including elements for defining the circumstances of observation, the methods of scoring, and some given conditions to be assumed. One item form for the electronics repair work might, for example, follow the form: "Given (diagnostic data) about a malfunction in (product), and given (conditions), candidate must (locate and replace) malfunctioning (part) with the work or response evaluated by (method of observation or scoring)." That particular item form may suggest many specific items. With all blanks filled, a change in any one of them defines the content of a new item. If the malfunctioning part is a condenser, replacing it may require soldering, and the solder connections can be evaluated either by an inspector's rating or by measuring the current flowing through the connection. Changes in diagnostic data or conditions can result in quite substantially different items, differing not only in content but in difficulty. Different item forms may require the candidate to develop diagnostic data, or to make judgments necessary for subsequent steps in a process, or simply to answer questions.

The number of possible item forms is obviously very large, even for a very simple job; more complex jobs might require, if not an infinite number of item forms, a prohibitively large number. However, a test content universe can be said to have been defined when the rules for generating item forms have been specified, even if no actual examples exist. The rules themselves, if fully stated, can identify the nature of responses that can be obtained and the variety of ways those responses can be evaluated.

DEFINITION OF TEST CONTENT DOMAIN

By establishing some item forms and rules for the generation of others, a test developer and his panel of experts will have shown the plausible limits within which testing can conceivably be done. It does not follow that all plausible item forms are worth developing. The panel of experts will probably dismiss some item forms out of hand as too costly or otherwise impractical. A process of elimination may be necessary to select from the universe of all possible observations a restricted subset for actual test development.

In one job analyzed by the writer, it was found that one task category consisted of reading material of varying complexity to obtain information fundamental to carrying out other duties of the job. The universe of admissible observations could be restricted to conventional reading comprehension testing. Facets included, in addition to the stimulus content material, facets of item format (multiple-choice, true-false, arrangement items, essay items, fill-in items, etc.) and facets for scoring responses (differential vs. unit weights, use of machine scoring vs. independent observers rating open-end responses, etc.). The operating decision was made, largely on the basis of the practical considerations of the number of people to be tested and the time period within which the testing had to be completed, to measure comprehension with conventional multiple-choice items following sample passages to be read. Rules for sampling material for the stimulus passages were established, and multiple-choice items -- a very conventional "item form" -- were chosen for development. The test content domain, therefore, consisted of passages to be read, sampled according to the rules established. and multiple-choice questions covering that material. The deficiency in the definition of the test content domain in that particular instance was that the rules for determining numbers of items, difficulties of items, and other related matters were never specified.

Test Specifications. In any case, with greater or lesser precision, with greater or lesser representativeness of the possible universe, with greater or lesser ambiguity, the panel of experts, under the guidance of testing specialists, will arrive ultimately at a set of specifications for test development. These will be in part content specifications, in part format specifications, in part response specifications, in part scoring specifications. There may also be some structural specifications for the inclusion of items; although the work sample test is intended to be a content-referenced test, the panel may specify limits of conventional item difficulty levels or discrimination indices.

TEST DEVELOPMENT

If the specifications call for a conventional job knowledge test, test items must be written. If the specified item forms require that something more than rote memory is to be invoked in testing for job knowledge, the items must be written to be challenging and to require thought while taking the examination. For hands-on performance tests, the elements (they, too, are items) of the tasks should also tap fundamental rather than superficial. The development of really good items is the foundation for either kind of work sample testing.

QUALIFICATIONS OF ITEM DEVELOPERS

Item writing or development is an art, and it is also hard and highly specialized work. Large test publishers or major civil service jurisdictions maintain on their staffs full-time, professional item writers; paradoxically, in highly specialized fields, the people most likely to have the necessary knowledge and full understanding of the implications of that knowledge are the people who have worked in that job, not those who have worked as professional item writers.

Wesman (1971) identified six points describing the combinations of abilities necessary to write good test items; they apply also to developing good exercises to fit performance test item forms:

1. The item developer must have full mastery of the subject matter of the test; full mastery does not imply merely acquantance with basic facts and principles but the understanding of them and their implications. This implies awareness of popular fallacies and misconceptions.

2. The item developer must have a clear understanding of the objectives of the test and of the reasons for testing. In educational testing, this implies an understanding of curricular and educational objectives as well as the specific test objectives. In developing a work sample test, it implies understanding of organizational values and of why doing certain things well on the job may be more important than other things that are relatively trivial but easier to put in a test.

3. The item developer must understand the characteristics of the people for whom the test is constructed. This means not only an awareness of the examinee's anxieties but also implies an awareness of the ignorance or clumsiness of potential examinees that can lead to mistakes, including acceptance of plausible wrong answers.

4. The item developer must be excellent in the use of language. This requires not only a useful vocabulary but skill in arranging words so that their precise meaning is inescapable. It applies as much to instructions as to verbal test items.

5. The item developer needs to understand specific techniques of item writing, including familiarity with different types of test items, their possibilities, and their limitations. Wesman points out that skill in item writing means more, however, than merely an understanding of item types. It requires imagination and ingenuity to create the kinds of situations, sometimes on paper, that will evoke expressions of knowledge. It requires similar imagination and ingenuity to abstract critical exercises from long and complex tasks.

6. Item developers who are skillful in one context or test type may find themselves less skillful in others; they need to learn their own skills and to collaborate well with others whose special strengths are complementary.

It seems clear that the combination of the professional item developer and the job incumbent is essential to the development of a job knowledge test. This combination is not likely to be found in any one person; test specialists must therefore train potential item writers in the skills of item development. A panel of experts is needed who have performed the job and know it well, either as incumbents or supervisors.

The first orientation meeting for panel members should clarify at the outset the purposes of the test to be developed and the general principles and values that guide the test construction enterprise. This should be followed immediately by a general instructional in item forms, item types, and item development. It is unlikely that imagination and ingenuity can be created in a brief training period, but they can certainly be stifled in such training if signs of them are not rewarded. A standard textbook, or, indeed, Wesman's chapter on item writing, can provide text material for such training; others (e.g., Boyd & Shimberg, 1971; Jones & Whittaker, 1975) can serve for examples of hands-on test items. As each item type is discussed, panel members should be encouraged to try to prepare the various kinds of items to fit the existing specifications, and examples of ingenious items should be brought to the attention of all.

Such training may seem largely superfluous if the test specifications have identified useful item forms and have specified in detail the test content. Even with such detailed specifications, however, item developers who understand their options well can be expected to be more creative in the invention of items within the item forms.

In short, item developers must be qualified, first, in terms of job experience and second, in terms of special training in the techniques of developing test items. A third qualification to be effective in developing items is motivation. The item developer must accept the purposes of measurement, believe in it, and have a desire to contribute to the development of an effective test.

ITEM POOLS

Although test specifications specify the number and kinds of items to be prepared and the content balance among them, the item developers should produce a surplus of potential items. Wherever possible, they should develop a pool of items at least 2 1/2 times the number needed for each item form. Such an item pool makes it possible to develop two randomly parallel forms of the test (Cronbach, 1971) without having to resort to items of questionable quality. Having two forms has a substantial number of benefits, not the least of which is the benefit in retesting people without compromising the security of the first form. However, the main reason for recommending two forms is that the evaluative data analysis to be recommended often requires them. The correlation of scores on these two forms, if nothing else, is an indication of the relevance of item content to the content domain and of the degree to which the content specifications of the test were clear enough to produce essentially similar instruments.

Job Knowledge Items. Job knowledge tests are not necessarily written tests. Comer (1971) described a trade test for the automobile service trade which used actual specimen parts from serviced automobiles as the source of items. The examinee could look at, handle, feel, or even smell each part. For each one, he was asked to (a) identify it, (b) describe its condition, (c) decide why it failed, and (d) decide whether it could be reused. The rationale is that an uninformed person would not know what to look for in examining a part. "He would not recognize pertinent symptoms, nor distinguish between normal, harmless scratches and discoloration, and true damage" (Comer, 1971, p. 50). For each part, the examinee filled in a blank with the name of the part and checked one alternative each describing condition, reason, and prognosis for further use.

The "items" in this test might be considered either the automotive parts examined or the questions asked about them. Since there were ten parts and four questions about each one, the test has either ten items or 40. For developing the item pool, it is useful to consider each question an item; for item analysis, there may be only ten items. If there are contingencies among the four questions in each part (e.g., if knowing the nature of the damage depended on proper identification of the part), then neither conventional methods of item analysis nor latent trait analysis can properly be used.

There are no similar uncertainties in most job knowledge tests; they typically are measured through the use of multiple-choice or other more or less objectively scorable, clearly independent test items. Nothing is needed here about suggestions for writing such items since there are many useful sources available (e.g., Ebel, 1972; Wesman, 1971). The item developer should be warned, however, that, although item writing looks easy, the only easy thing about it is an easy superficiality.

The problem is reflected clearly in the study of occupational licensing practices by Shimberg et al. (1973). They attributed the many flaws in written licensing tests to four categories:

1. Few of the licensing boards did any planning for the test

beyond a vague outline, and some did not even have an outline.

2. Many state and local boards have a compulsive preference for essay tests; essay items are often ambiguous and scoring procedures are usually unreliable.

3. Where multiple-choice items are used, many licensing boards prepare items which are answered correctly most readily by people who have memorized the review books from which they were drawn.

4. There is little evidence of any concern for statistical item analysis or any other item evaluation procedures; there seems to be an attitude that once the test items are written, the problems of test construction are over.

Items in the Total Job Test. Jones and Whittaker (1975) refer to "total job" tests. In these, the examinee is simply doing the job (or a thorough replica or simulation of it); the only thing that identifies it as test performance is that the work is done under standardized conditions, that the performance is systematically observed, and that the observed performance is evaluated directly. What are the "items" to be developed for such a test?

Perhaps the ultimate "total job" test is a carefully designed probationary period in which the worker is systematically rotated among different assignments or work stations. Each stop in the rotation may be an item or, alternatively, a subtest in which the items might be units of time, such as production or scrap per hour, or points on an observer's or inspector's checklist.

This is a limited form of testing. Only relatively routine, short-cycle jobs fit easily into this frame. For most jobs, samples of the work to be done are practical necessities. Such samples can for convenience be divided into those in which the sampling is rather direct and obvious and those in which the sampling evolves from a process of abstraction resulting in a test that is quite different in appearance from the actual job.

Items in Direct Work Samples. The item forms established from test specifications determine the nature of the items. Each item form identifies a task to be performed and the circumstances in which it is to be performed; that task is an item. As in Comer's (1971) test for the automobile service trade, however, there is always some uncertainty about whether it is the task as a whole or the component scorable parts of it that are the literal analogs of test items. The solution to the uncertainty, as before, seems to be in determining whether there are contingencies in the finer divisions of scoring performance that will interfere with effective item analysis.

Foley (1977) speaks of "scorable products"; Maier, Young, and Hirshfield (1976) refer to "scorable units." In a literal sense, any element in performance, product, or observation that can be graded, rated, classified, or otherwise scored is a de facto item. To identify scorable units, the panel of experts may need to develop successively smaller divisions of tests, subtests, item frameworks, and independently scorable units. It must determine precisely what events or activities or attributes of products can be observed and how they can be evaluated.

A few examples will identify the variety of work sample items that can be invented:

1. In marksmanship performance tests, firing a specified number of rounds from a specified position a specified distance from the target is an item (Wright & Mead, 1978).
2. In a dental hygienist's work sample, performance is observed

by an observer with a checklist of specific steps in each of several task sequences; each step is an item rated by the observer (Boyd & Shimberg, 1971).

3. In an offset printing work sample, items are statements in lists of violations of good offset press operation and violations of safety rules; each violation noted is checked (Boyd & Shimberg, 1971).

4. In a test for truck mechanics, one part of the battery is the discrete task, "remove and replace clutch plate." Nineteen penalty scales for specific examples of poor performance, under five headings, were rated for degree by an observer (Jones & Whittaker, 1975).

5. A metal lathe operator may be asked to make a taper plug (not because it is an actual piece, but because it requires so many kinds of turning) according to engineering specifications. Items include (usually dichotomously) the match of the product to eleven specified dimensions and could include ratings of both smooth and knurled surfaces (Jones & Whittaker, 1975).

6. In a tank crew gunnery test "engagements" are items; the 28 engagements varied according to gun used, time of day, responsible crew member, mode of firing, target type, target range, and other facets of the exercise. Items were scored as hits or misses, either on first trial or on second trial (Wheaton et al., 1977).

7. The REALTRAIN battle simulation has no clearly identified a priori items; however, a posteriori items can be derived from a net control station data recording sheet identifying casualties claimed and confirmed (Shriver et al., 1975). It is worth noting that the exercise in its entirety, as are many other simulations, actually constitutes a single item.

8. In a performance test for radar electronic maintenance technicians, 81 problems involving use of test equipment, adjustment or alignment of test equipment, and many others led to a total of 133 "scorable products" to be completed satisfactorily (Foley, 1977).

Most work sample tests are reported with no clear identification of items or scorable units of measurement. Apparently, test developers have given little attention to the concept of an item of measurement. Many large-scale, single-item simulation exercises, carefully and elaborately planned, are excellent management or training tools -- like REALTRAIN, which was designed explicitly for training purposes -- without being particularly good measures of anything. In nearly any measurement by testing, items of measurement are important in one of two major roles: either as indicators of units on a fundamental scale of measurement or as replications, as in the several items of an achievement test. In many work samples, the scored items are contingent items; "passing" the item depends upon prior success with a previous item. Such items are certainly not replicates; neither will they form a scale without very careful planning. Treating overall performance on an exercise as a whole, or as a single item, risks seriously unreliable measurement.

It is here asserted that better measurement will be obtained from direct work samples only when item developers concentrate on identifying and developing work sample analogs of independent test items.

Items in Abstracted Work Samples. The distinction between a direct and an abstracted work sample is no more than the distinction between ranges on a common continuum; they are not distinctly different kinds of tests. The difference is that direct work samples have at least the appearance of "real" work assignments. They do, of course, involve some abstractions; the machinist who is required to make a taper plug, for example, may never again have a specific task assignment which requires all of the specific kinds of cutting required by the test -- but assignments from a variety of "real" jobs are abstracted from them and put together in one "artificial" task assignment.

Abstracted work samples may seem less like the real job than like a test, but they nevertheless clearly sample the activities and the skills of real jobs. An example is the typical In-basket Test. An In-basket usually involves inventing a make-believe organization with imagined organizational relationships, problems, and required decisions. Within this imaginary framework, pieces of paper are created to be similar to papers found in actual in-baskets of real people in real organizations. The examinee must pretend that he is really in the make-believe world of the imaginary organization, that he is new to the organization, and that he is working at night with no one else around to give him needed information not in his packet of test materials. Yet the persistent reaction of people who take such tests is that they are "realistic" -- that is, the In-basket Test, if reasonably well constructed, poses such real-to-life problems that its artificial aspects fade into insignificance as the examinee becomes more engrossed in its basic reality.

Abstractions may maintain the "hands on" character of the direct work sample while seeking independently observable performances as items. To illustrate the problem, assume the home workshop project of making a lamp out of a bottle set on a specially turned wooden base. Four independently acquired skills are used: turning the base, cutting the bottom off the bottle, assembling the various parts, and attaching the wiring. A direct work sample would give the examinee the necessary plans, materials, and tools and set him to work. The results might be scored in terms of whether the lamp works, whether it is steady on the table, or ratings of various aspects of its appearance. An abstracted work sample, in contrast, might be four quite different tests. One might be a turning test, in which several wood turning skills could be demonstrated without designating any product as a lamp base, the second might be a similar test for bottle cutting, the third an assembly test in which a previously turned base and bottomless bottle were provided along with the other lamp parts, and the fourth the simple wiring of an assembled lamp. At no point is one step of the lamp-building procedure dependent on how well the previous steps had turned out, nor is it necessary that the worker have a lamp in mind as the final product while taking the first two tests. All that is necessary is that the component tests (or subtests) tap the necessary skills.

Abstractions may also minimize or even abandon the "hands on" feature. Job knowledge tests are usually abstract work samples. That is, the knowledge components of the job content domain have been abstracted to form the nucleus of the test content universe and domain. The result is in no sense a full work sample, but it can be considered a partial work sample -- a sample of what expert judges considered an essential component of the total job. Comer's test for the automobile service trade is an example. A direct work sample could have been developed that would require the demonstration of skill in removing the parts central to the test, diagnosing them, and taking appropriate corrective action. The essence of corrective action, according to Comer (1971) is the knowledge one brings to and derives from examining the part and deciding why it might look as it does.

What is here called an abstract work sample may be very much like the test Foley (1977) termed a symbolic substitute for a performance measure. What is intended by the term is, however, more than symbolism. What is intended is literally pulling from the whole an essential element -- an abstraction -- on which the abstracted performance can be used to infer performance on the whole.

The process of inference clearly identifies the abstraction as a test; it places the focus on the score and on the necessity for an empirical validation of the inference. It also places focus on item definition as the process of test construction; the item form is seriously taken as a prescription for an item, e.g., "Given all necessary parts and ordinary hand tools, candidate assembles lamp ready for wiring."

Requirements for Scorable Items. There are many kinds of work sample tests. A multiple-choice job knowledge test is an abstract sample of the job (Campbell et al., 1973). So is an abstraction of decision processes (NCARB, 1976) or of manipulative activities (Comer, 1971) -- or, for that matter, a standardized observation of a full cycle of actual job performance for a specified time period. Whatever the nature of the sample, it can usually be analyzed into component parts, and performance on these components can usually be scored, graded, or rated; each component is therefore a test item which, aggregated according to a specific rule, yields an overall score or subscore. These items may be as complex as a battle simulation or as simple as a brief time span.

What is required for such items during the development of an item pool is that each be independently scorable and that each be traceable to its origin in the job content domain.

ITEM ANALYSIS

Once a pool of items has been prepared, item analysis is necessary both to identify poor items and to verify the match of actual items written to the defined test content domain. Verification requires the systematic collection of judgments; other kinds of poor items can be identified by conventional item analysis.

Retranslations. The judgments concerning the fit of items to the test content domain can be obtained in a variety of ways. One possible method is an adaptation of the "retranslation of expectations" proposed by Smith and Kendall (1963) for the development of behaviorally anchored rating scales.

In brief, the first step in the Smith-Kendall procedures requires a group of experts to identify the dimensions along which ratings of performance should be made. An analogy to content sampling is the identification of components of the job content domain. A second step is to write behavioral statements for different levels of each dimension, for which the parallel step in test construction is writing test items. A third step convenes an independent group of judges to allocate each behavioral statement to the dimension from the original list which it fits best. The "retranslation" analogy is the practice of translating a passage in English into a foreign language, and then retranslating the foreign language passage back into English. If the two English versions -- the original and the retranslated -- match, it is assumed that the foreign language translation was satisfactory. By the same analogy, if items are developed to fit one or more components of a defined test content domain, then an independent panel of judges should be able to allocate each item to its proper components. If there is no consensus about the fit of an item in the content domain, it is identified as a poor item; where there is consensus, the appropriateness or relevance of the item for its intended purpose in the domain is verified or established. If there are components of the overall content domain for which the items written have not "retranslated", then deficiencies in the item pool are identified.

Content Validity Ratio. Lawshe (1975) presented an equation for computing a statistic describing the degree to which an independent panel of knowledgeable experts considers the knowledge or skill measured by an item to be essential to the performance of the job. He called it the content validity ratio. For a variety of reasons, including the fact that the technique does not identify deficiencies in content sampling, this writer prefers to call it a job relatedness ratio; it provides an index number for evaluating the relevance of an individual item for job performance.

The job relatedness ratio is computed by the formula:

$$jrr = \frac{n_e - \frac{N}{2}}{\frac{N}{2}}$$

in which n_e is the number of panelists classifying the content of the item as essential to satisfactory job performance (as opposed to "useful but not essential" or "not necessary"), and N is the total number of panelists. The ratio is a direct linear transformation from the proportion n_e/N, but it has certain advantages other than simple proportion. If less than half of the judges consider an item essential, the job relatedness ratio is negative. If exactly half say that the knowledge measured is essential, the ratio is 0. If all say that the knowledge is essential, the ratio is 1.00. In short, the range of the job relatedness ratio corresponds to that of a correlation coefficient and may therefore be interpreted in fairly familiar ways. Another advantage is that the minimum values for the job relatedness ratio to be statistically significant at different panel sizes have been reported by Lawshe.

The ratio does not identify components of a test content domain omitted from the item pool judged. This seems to be a relatively trivial matter. If the job analysis defining the job content universe and domain has been thoroughly carried out, and if the judgments extending the job content domain to the test content domain and to test specifications have been thorough, then item development to meet those specifications should provide a rather complete coverage of the job domain. As a matter of fact, any component of the domain that systematically loses items on the basis of low job relatedness ratios probably should not have been in that domain in the first place. In such a case, the orginal panel of experts should certainly give serious consideration to redefining the domain without that component or to writing new items that might have a better chance of being judged essential.

Thus the job relatedness ratio offers real advantages in item analysis. It provides a procedure for systematizing and documenting the judgments of a panel, it provides an index number that can be readily interpreted, and it provides a record of the judged importance of individual items that can serve as evidence of the job relevance of individual items in the case of litigation.

Index of Item-Objective Congruence. A combination of procedures like those described above characterizes an overall statistic reported by Rovinelli and Hambleton (1977). To place their index in the context of personnel testing, the reference to curricular objectives may be taken either as a reference to component tasks in defining a job content domain or universe or to the task category defining the job content domain.

Each judge in a panel evaluates each item in the pool for each of the components. Evaluation is expressed on a 3-point scale of +1, 0, or -1, indicating either a clear allocation of the item to the component, indecision, or a clear belief that the item does not fit that component.

The computation of the index is based on a data matrix in which the judgments of the individual content specialists may be considered the columns and the objectives or job components may be considered the rows. The index of item-objective congruence is then given by the equation:

$$I = \frac{(N-1)\Sigma X - \Sigma\Sigma X + \Sigma X}{2(N-1)n}$$

where

I is the index of item-objective congruence for a specific item on a specific objective for a component,

N is the number of objectives for components,

n is the number of judges,

ΣX is the sum of the ratings for that item by the judges, and

$\Sigma\Sigma X$ is the sum of the summed ratings across objectives.

The resulting index, like the job relatedness ratio described above, is an index number that conveniently ranges between -1 and +1 with the 0 point indicating that all judges are undecided about the allocation of the item. The major feature of this index is that it is not simply a global judgment of relevance, but is a judgment of relevance to specific components of the content domain. Rovinelli and Hambleton recommend that the index be computed for every item for every objective, a recommendation which has the merit of identifying ambiguously assigned or redundant items. Like Lawshe, they make no particular recommendation concerning a minimum value of the index deemed acceptable for item inclusion; this, they say, will depend on experience both with a panel of content specialists and with the use of the index. A disadvantage of this index in relation to the job relatedness index reported by Lawshe is the absence of a significance test; this is probably not a serious defect; one could easily be developed.

Conventional Item Analysis. The statistical item analysis techniques conventionally used in norm-referenced test construction may also be applied to items in work samples. Whereas the foregoing analyses have been based on a priori judgments of the items, conventional item statistics are based on actual responses to the test items from a specific sample of examinees. The sample should be as representative as possible of the population with which the test is to be used. If the test is to be used for certifying knowledge or proficiency on a job (or denying such certification), then the sample should consist of candidates for certification. This conventional requirement for a representative sample poses some special problems for the item analysis of work sample tests. People applying for certification are probably highly self-selected so that most of the candidates will pass most of the items; that is, there may be little variance either in item responses or in total test scores. Conventional item analysis techniques require at least some variance in both.

The most useable item statistic is its difficulty or easiness level. It is conventionally computed simply as the proportion of people in the sample passing the item (e.g., answering a question correctly). It used to be said that the average difficulty level should be somewhere around .5, a prescription with little value for the development of a content-referenced work sample test. One might determine the rank order of items in the item pool in terms of their difficulty; the most difficult items and the easiest items can be reexamined for the appropriateness of their inclusion. An item for a critical content component might, if excessively easy, fail to measure adequately the salient knowledge or skill. On the other hand, an item that is too difficult might also fail to measure it because of ambiguity, fuzzy instructions, or some other defect. The computation of the item statistic, therefore, should not be considered the final evaluation of the item. The panel of experts can be reconvened to reconsider items in the light of their descriptive statistics, and any item which, in its collective judgment, is either too easy or too difficult, may be revised.

The other conventional item statistic is the item discrimination index. If an item correlates well with the total score, it is assumed to differentiate well between those who are knowledgeable or skillful and those who are not. A low correlation identifies items assumed to be poor in distinguishing the certifiable from those who are not.

The absolute value of these correlations is usually trivial since it depends so much on the variance in the available sample. However, the item-total correlations can be used for rank ordering the items.

High item-total correlations are usually considered unnecessary in content-referenced tests; such tests need not have high levels of homogeneity (Cronbach, 1971). However, such tests should have at least some functional unity. Functional unity means that items within a test should somehow hang together; everything in the test should be at least a little bit correlated with everything else. In other words, the item-total correlation need not be high, but it should be positive. Zero or negative correlations identify either poor items that should be revised or replaced or subsets of items that should be independently scored. It is poor measurement when a test score represents an unknowable combination of independent or even negatively related attributes.

SCORING AND CALIBRATION

The conventional method of scoring a conventional test is simply to count the number of items with correct or passing responses. A multiple-choice test may be scored according to a formula, such as the number of items correctly answered minus a fraction of the number of items answered incorrectly (the fraction depending on the number of optional responses available). Each item is a unit. One item, easy or hard, counts just as much as any other item; differential weighting of items is traditionally considered useless with large numbers of items. Hambleton et al. (1978) offered five methods for estimating examinee domain scores or "true proportion correct scores."

Item analysis may, particularly with small numbers of items, provide a basis for differential weights, but it is rarely if ever done. Differential weighting of items in work sample tests is rather common, but it is usually based on a priori judgments of relative importance rather than on item statistics.

Strictly speaking, it is usually improper to refer to a conventional score on a psychological test as a <u>measure</u> of an attribute; it is better to describe it as a sign or <u>reflection</u> of the attribute. Whether the score really can be used as an indicator of the aptitude is a question to be answered as its construct validity.

It has been pointed out that concepts of validity are unnecessary in fundamental measurement. If work sample tests are to be more than reflections of level of performance, if they are in fact to be formal measures of performance variables, then scoring should be more than merely a count of right answers. The score must form a scale with fundamental mathematical properties, at least the property of ordinal transitivity. The translation of scores into such scales involves item or score calibration. One simple method of calibration will be proposed, based on the logic of what Ebel (1962) called the content standard score. Another one will be briefly mentioned (Anderson, 1976). The major discussion will, however, be devoted to latent trait theory, a well-developed approach to calibration through so-called "sample-free" item analysis.

LATENT TRAIT ANALYSIS

Latent trait theory overcomes the problem of over-reliance on a specific sample. Wright (1968) has demonstrated that calibrating items using the one-parameter Rasch model will yield ability estimates in new groups of people that are independent of the ability distribution in the particular sample of people on whom the test was calibrated. Moreover, estimates of item difficulty level can be made more or less independently of the distributions of difficulty in the set of items used in making the estimates. That is, the item and ability parameters are essentially invariant across samples of people and of items; this, according to Wright, is the ultimate in objectivity in measurement.

The Rasch model estimates only one parameter of the item characteristic curves. This is generally justified on the grounds that three-parameter models are less successful in finding invariance for parameters other than the difficulty parameter, although Slinde and Linn (1978), finding insufficient evidence of invariance using the Rasch model thought additional parameters would help. Item discrimination values are less reliably estimated, and the estimation of so-called "guessing parameters" is often seriously unreliable. Rudner (1977) has demonstrated similar objectivity using a three-parameter model, but the estimates of the a and c parameters were considerably less reliable than those of the b (difficulty) parameter.

Proponents of different latent trait models seem adamant in their preferences; these reports will not take sides in disputes on the general relative merits of the different models. However, for the purpose of developing a job knowledge test, the application to be discussed first, the three-parameter model will be used as the prototype. There are two reasons. First, it seems to be the more general model, giving more information than is possible with the Rasch model. Second, it provides another form of analysis by which one kind of "poor item" may be identified.

Consider Figure 1. It proposes to show three-parameter item characteristic curves for four items in a given job knowledge test. The items are arranged in increasing order of difficulty so that they may be said to form a genuine scale, at least p_b probability of a correct response. The a parameter, the discrimination parameter, is essentially the same for items 1, 2, and 3. It is markedly different for item 4. For people with high levels of ability, such as θ_2, items 1, 2, 3, and 4 form a progression of difficulty with item 1 the easiest (highest probability of a correct response) and item 4 the most difficult. For people with levels of ability at θ_1, however, item 4 is the easiest, that is, it has the highest probability of a correct response.

Alternatively, the problem can be seen by examining the ability level required for a specific probability of a correct response. At p_b, the ability level requirement increases from item 1 through item 4. At p_a, however, the ordinal positions of the items in terms of required ability level is 1, 4, 2, 3. From the point of view of

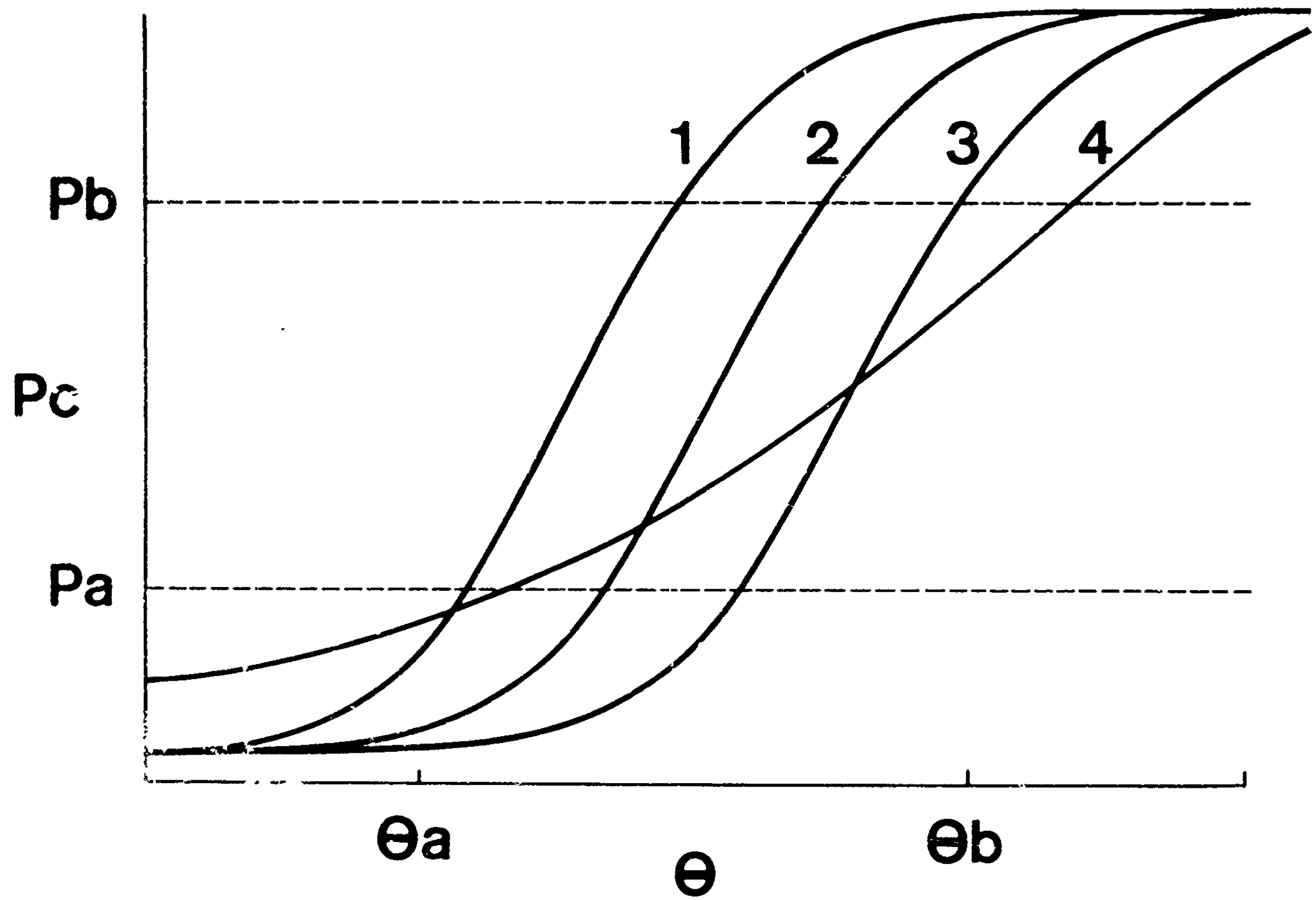

Figure 1. Three-parameter item characteristic curves, for four hypothetical items, showing changing order of difficulty at different ability levels.

formal measurement, the first three items are excellent; they form a true, transitive scale of measurement regardless of who is tested. Item 4 spoils the scale; with it included, the transitivity of item order varies according to ability level.

Clearly, item analysis by latent trait item characteristic curves can not only lead to item calibration but also to identifying poor items from the standpoint of the formal properties of measurement.

Suppose that isolated examples of such intransitivity were found. These would identify items that should either be deleted or revised. Suppose, however, that a clustering of items with similar discrimination values were found and that they were associated with different components of the content domain. Such a finding would be a reason for establishing independent scales for measuring independent components, even though the scales might appear to have some degree of functional unity.

Four procedural matters or assumptions need to be considered in applying latent trait theory. The first is the assumption of a single dimension. There are multi-dimensional methods of latent trait analysis, but the models most likely to be used, particularly in job knowledge testing, are those that assume a single underlying trait. The assumption of unidimensionality does not pose a rigorous demand, but the items to be scaled should possess a reasonable degree of homogeneity. This is somewhat at odds with the more general statements in these reports and elsewhere that high levels of internal consistency are not absolutely essential in content-referenced tests. The argument has been that any score on a work sample test should represent some functional unity, with the elimination of any items that have 0 or negative correlations with the composite. For latent trait analysis, however, absence of orthogonality is not enough; substantial levels of internal consistency are required. For a latent trait analysis, therefore, a factor or cluster analysis is needed so that independent scores can be obtained for each of the component dimensions. There will then be as many latent trait analyses as there are dimensions in the item pool.

The second assumption is the assumption of local independence. In general, local independence means that performance on one item is not dependent on performance on other items. The most obvious violation of the assumption occurs when there are contingent items; for example, if item 2 can be answered correctly only if item 1 has been answered correctly (as in certain arithmetical reasoning sequences and work samples), then the assumption of local independence has clearly been violated. Other violations of the assumption are less obvious, and there is no convenient test for violations. However, strong evidence of unidimensionality is generally accepted as evidence that the requirement of local independence has also been satisfied.

The third point is a procedural matter. The LOGIST computer program for the three-parameter model available from the Educational Testing Service (Wood & Lord, 1976) requires very large sample sizes, calling for 50 items and 2000 subjects as a desirable minimum for determining item characteristic curves. This requirement may, perhaps more than anything else, be responsible for the early lack of interest in latent trait theory among applied measurement specialists. New programs have made it possible to estimate item parameters reliably on substantially smaller samples. Bejar (1977), reported results of sex differences in item characteristic curves using 178 males and 143 females in independent analyses of 20-item scales. With the simpler Rasch model, Wright and Stone (1978) reported a study with less than 20 items and 40 people. Despite uncertainty about the statistical power of estimates based on small numbers of people or of items, it seems clear that progress in computer programming is resulting in more efficient algorithms requiring less luxury in the size of either sample. Latent trait analysis is becoming practical.

Nearly all latent trait programs now in use are iterative programs. That is, they begin with the raw score as the first estimate of the underlying ability and then estimate item parameters. A second ability estimate is based on these parameters, and the whole process of estimation enters a second iteration. The procedure continues until the solution converges, that is, until successive iterations produce little or no change in the item and ability parameters. Some sets of data simply do not converge. Where this happens, it indicates that the data will not fit the model being used and, perhaps, an alternative model might be attempted. It is more likely, however, where convergence fails that it is less a problem with the model than a problem with the data fitting the assumptions of the model. The assumption of unidimensionality deserves particular attention. One or two aberrant items in an item pool that clearly do not fit the underlying dimension can be responsible for difficulties in achieving convergence. The problem may disappear when they are deleted.

The fourth point is the need to check the fit of the data to the model. The question of the fit of the model has two components. One is the degree to which the theoretical item characteristic curve will in fact fit the data. A normal ogive or logistic curve simply may not be acceptable fits for an item; checking this out is a simple curve fitting problem. The other component is the fit of people to the model. Some people may not respond as the model indicates that they should. If the set of responses to the items on a test from a person of average ability follows the model reasonably well, he will get most of the easy items correct, will tend not to get so many intermediate items correct, and will not succeed in giving correct answers to most of the difficult items. If a person responds carelessly, however, there may be little in the proportion of correct responses across different difficulties. If the person is using a test-taking strategy that capitalizes in some unusual way on certain item characteristics, it may be easier for that person to get correct responses to some generally difficult items for which that strategy is especially useful than to get correct answers for easier items where the particular trick does not work.

Thus, both items and people may be identified which do not fit the model, and it may be necessary to delete aberrant items or people or both to achieve a satisfactory solution.

Other Latent Trait Models. Although most work in latent trait theory has been done with conventional paper-and-pencil test items, usually multiple-choice, that can be scored dichotomously, latent trait models can also be applied to observations of work sample performance or product. Some of the "items" in a work sample will, in fact, be dichotomously scored. An example is whether a mechanic disconnects battery cables in the right or wrong sequence. Other observations might be continuously scored, for example, the actual deviation of precise measurement from a specified measurement in drilling a hole. The latter item, of course, could also be scored dichotomously as either within or exceeding allowable tolerances.

By using a free response model (Samejima, 1973), one can incorporate into a single latent trait analysis all observations fitting a particular dimension regardless of the scoring format. The advantage is that dichotomies are not necessary; a test developer can take special advantage of the added information included in continuous responses.

Although multidimensional latent trait models have been proposed, no work with any of them is known to the present writer. Nevertheless, if data are not unidimensional, one could certainly submit them to factor analysis or cluster analysis. The Bejar (1977) study applied the Samejima model to data both multidimensional and reported with continuous responses.

Scoring Procedures. Suppose that three-parameter item characteristic curves have been computed for each item of a test and that they differ substantially on all three parameters, particularly on the difficulty parameter. For low ability examinees, correct responses to difficult items occur primarily by chance. Their responses to these items are not very informative; in fact, information curves for items with high difficulty levels are very nearly at zero for low levels of ability, peaking with substantial levels of information only in the high ability ranges. Obviously, then, the traditional score of the number of items answered correctly is not as informative for low-ability people as the score obtained by ignoring (i.e., assigning a zero weight to) the very difficult items (Lord, 1968).

An optimum weight curve can be computed from the item characteristics curve to show the optimum weight to be assigned to each item, at each difficulty level, in scoring the test. Lord noted three facts about the optimum weight curve:

1. As ability increases, the optimum weight increases, the increase gradually becoming nearly horizontal, asymptotic to a value proportional to the discrimination parameter of the item characteristic curve. At high ability levels, then, optimum item weights depend on item characteristics, not on ability level.

2. Optimum weights are lower for lower levels of ability because of the increasing effect of random guessing.

3. At very low levels of ability, the optimum weights for difficult items become very close to zero.

There is a bothersome paradox in all of this: one must know the ability level one is trying to measure before one can determine optimal weights for measuring it! If scoring can be done by computer, a maximum likelihood estimate of the ability parameter can be estimated for each person (Lord, 1977). Alternatively, one can use, by computing the maximum likelihood estimator, the conventional raw score as a preliminary estimate of ability. This value can be used in charts of optimum weight curves to find a nearly optimal set of weights for a particular examinee. The test paper can then be scored to obtain a nearly optimal total score and, if needed, a conversion table can be established for converting these "nearly optimal" scores to standard score scales of ability. (Lord, 1977, offered a variation of this procedure.) The fact that the optimum weight of an item is independent of the weights of other items is what makes tailored testing practical; the score can be expressed along the same scale even if different items have been used in arriving at it. There are alternative scoring options. From the first fact noted above, it follows that the discrimination parameters of the item characteristic curves can be used as weights, without regard to examinee ability level, if measurement focuses primarily on the upper ability levels. A slightly different value can be derived for multiple-choice items where there is guessing:

$$w = \frac{a}{1 - c} \cdot \frac{p - c}{p}$$

where

w is the weight to be assigned to an item,

p is the conventional item statistic, the proportion giving the correct answer to that item,

a is the discrimination parameter of the item characteristic curve, and

c is the lower asymptote (the "guessing" parameter) of the item characteristic curve (Lord, 1977).

Scoring systems can also be based on item difficulty parameters. Considering items as if they were arranged in slightly increasing increments in the order of difficulty, and assuming that guessing does not occur, an individual's score on the test can be defined as the difficulty parameter of the next item just after the last item correctly answered. Another option is to use the average difficulty parameters of the items answered correctly; under certain conditions in tailored testing, it is essentially equivalent to the maximum likelihood estimator (Lord, 1974).

In certification testing, the importance of the choice of scoring method depends on the level of ability at which certification is to be granted. The choice of an optimal scoring procedure is critical if the decisions are to be based on minimal competencies which are relatively quite low. If competency implies a high level of ability, however, no great infomation loss occurs even if the simple number of right answers is used as the score. At these levels, the only advantage in using the parameters of the item characteristic curves as the basis for scoring is that they provide a standard scale independent of the particular distribution of cases or particular set of items used in constructing it.

The standard scale is especially valuable in work sample testing. There are often missing data; either the observers disagree on what they have seen to the point where no consensus is possible, or they miss seeing something important, or the performance follows an atypical pattern. Latent trait scoring removes many of the problems these events pose for traditional scoring.

The above discussion has centered on the three-parameter model; scoring by the Rasch model is somewhat simpler, although it is also a maximum likelihood procedure calling for computer estimation for maximum precision. The only item parameter, of course, is the difficulty level. Scores may be expressed in terms of difficulty levels, as before, or in terms of units of measurement (Wright, 1977).

<u>Advantages of Latent Trait Analysis</u>. A latent trait analysis has as its principal advantage the fact that it offers genuinely formal measurement of a mental trait. That is, the principle of transitivity and the principle of additivity are both assured through the use of latent trait analysis. A second advantage in scoring tests by latent trait analysis is the increased precision in measurement. Since a work sample test is developed to make decisions about individuals, maximum precision in measurement is an ethical obligation.

For widespread operations, such as military qualifications testing, these facts (and its basic nature) make latent trait analysis independent of the idiosyncracies of time or place. That is, the score obtained by one candidate for promotion at one time in one location can be interpreted on the same scale, and with the same precision, as the score of a different applicant obtained in a different physical location at a different period of time. It follows that promotional policies can be far more uniformly administered than through measurement techniques that place heavy reliance on samples of people for determining the scores.

An almost equally important advantage of the latent trait analysis, particularly for job knowledge tests, is that it provides an item pool which can be dipped into for a small set of items useful for measuring precisely the knowledge level of an individual candidate and a different set of items equally useful for a different candidate. Their scores will be on an underlying scale common to both examinees even with non-overlapping sets of items. For this reason, test security is less a problem with conventional job knowledge tests.

Contemporary society is greatly concerned about possible racial and sex bias in measurement; the latent trait analysis makes it possible to identify race-by-item interactions or sex-by-item interactions and to delete items contributing to such interactions from the item pool (Ironson, 1977). This is not to be confused with main effects due to race or sex, which may represent true differences between the groups, but it does identify where the probability of getting a correct response at a given ability level varies with race or sex.

Finally, latent trait analysis provides a firmer foundation for evaluations of possible adverse impact for minority groups or women. According to current Federal Executive Agency Guidelines (1977), a test is said to have adverse impact on a group when its rate of selection is less than 80% the selection rate of the subgroup with the highest percentage selected. The existing Guidelines of the Equal Employment Opportunity Commission (1970) require a test user to choose the test with the lesser adverse impact when there are two or more possibilities of similar validity. It is not entirely certain that the level of validity is a governing issue; from the point of view of fair employment practices, there is a substantial social and governmental pressure to seek tests with lesser adverse impact, irrespective of their validities.

One response to this has been to look toward work sample testing as a possible approach to employee selection with lesser adverse impact than ordinary paper-and-pencil tests have. Schmidt, Greenthal, Berner, Hunter, & Seaton (1977) recently published a study in which they demonstrated less adverse impact for a job sample than for a written aptitude test.

A methodological question might limit the generalizability of their findings.* Figure 2 identifies two tests, each with a different test characteristic curve. The test characteristic curves show differences in the discrimination value and essentially equal difficulty levels. That is, the slope on Test 1 is steeper than that for Test 2. Assume true ability differences on the underlying latent trait indicated by A and B on the abscissa. If these true differences exist, they will be obscured if Test 2 is used and exaggerated if Test 1 is used; that is, relative to the true ability scale, the obtained differences between the two groups on Test 1 is greater than the true difference, while the obtained difference on Test 2 is less than the true difference. In either case, the obtained score difference distorts the true difference between the groups. A similar figure can

*The author, with Gail Ironson as senior author, is preparing a detailed critique of the article by Schmidt et al.; it is sufficient here simply to illustrate the nature of the argument.

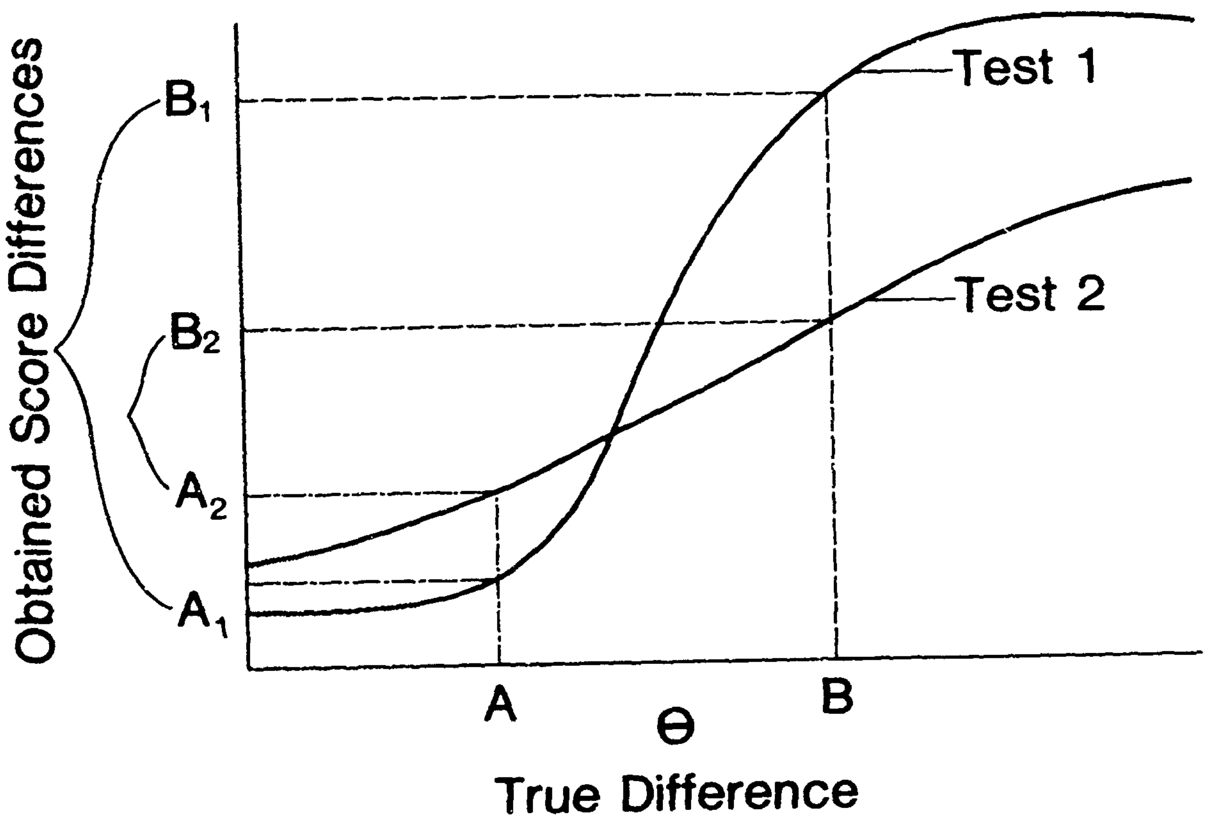

Figure 2. Obtained score differences between groups as a function of different test characteristic curve parameters

be drawn to indicate similar distortion in group differences in obtained scores where the difficulty levels vary and the discrimination values are essentially similar. Where the latent trait analysis of job knowledge tests and of any other tests with which it is to be compared, observed score differences can be interpreted in terms of the underlying latent ability so that different tests measuring the same ability can be compared on a common metric.

Possible Alternatives to Latent Trait Theory. The essential feature of latent trait theory is that it provides a content-referenced scale for interpreting scores and that it is not dependent on the distributional characteristics of specific samples. It provides a formal rather than the traditional psychometric form of measurement. The question arises whether there are alternatives to latent trait analysis to accomplish similar ends.

One possible alternative is the functional measurement advocated by Anderson (1976). It will not be discussed in this paper since it has not yet been applied to standardized tests. It is, however, a technique which develops a formal scale of measurement that is not dependent on the characteristics of individuals chosen for scale development; it is based upon theoretical mathematical functions.

One alternative is to interpret total test scores in terms of a score on a smaller subset of items with formal psychometric properties. The idea emerges, first, from the content standard scores proposed by Ebel (1962) and, second, from the writer's insistence on functional unity in content-referenced tests. It may be recalled that Ebel selected ten arithmetic items, each representing a different kind of arithmetic operation. No scale was identified for ordering these items. Scores on the total arithmetic test were interpreted in terms of the number of items probably correct in this standard set of ten.

Suppose the ten items were scaled on some dimension. They might be scaled in terms of difficulty, the level of knowledge required, the social importance of the arithmetic skill involved, or some other dimension of interest. The scaling can be done using the method of equal-appearing intervals, and all items in the total test can be scaled. A small subset of items can be chosen to meet the requirements of a Guttman scale where the discriminal dispersions of the judgments do not overlap. The resulting items can be ordinally ranked, differences between items can be expressed in a common metric, and the metric can be expressed as the specific dimension of interest (difficulty, importance, etc.). The items form a content standard scale with known psychometric properties.

The subset of items can be used as a total test; however, it is generally believed, in keeping with the Spearman-Brown function, that a small set of items gives substantially less reliable scores than can be obtained from a larger set. Therefore, the conventional score on the total test from which the standard scale is drawn is a more reliable score. It can, however, be interpreted in terms of the standard scale by the simple expedient of developing a regression equation for standard scale values from total test scores.

The principle and procedures can be applied to any work sample where expert judges can form a reasonably reproducible scale along a dimension of interest. If the dimension of interest is difficulty, conventional item statistics can be computed as the scale value, and the "discriminal dispersion" can be expressed in terms of the conventional standard error, $1/\sqrt{pq}$. Brenner (1959) demonstrated that tests could be developed by these standard item analysis methods to yield satisfactory reproducibility coefficients.

CUTTING SCORES

Administrative convenience seems to demand that qualification testing be done with fixed standards of mastery designated by rigid scores, above which examinees are classified as masters, and below which they are classified as non-masters. There is no way to determine a cutting score empirically in the absence of assumptions or judgments. Shepard (1977) put it well: "Performance standards do not inhere in nature; they have to be decided upon by fallible people."

The use of a mastery cutoff point is, perhaps, inevitable, but the actual scoring of tests should be on a more refined scale. If the purpose of the cutting score is the classification of examinees, it should be recognized that the mere fact of misclassification is not the essential error. A serious error, perhaps the most essential error, is the <u>degree</u> of misclassification. Unless scores vary along a continuum, no evidence can be deduced for determining the degree of classification error. Certainly, if the only errors of classification are minor ones, the "close calls," e.g., the erroneous classification of people as masters when in fact they are almost masters, they are relatively minor. A truly serious error of classification is when an individual who is a true master, substantially above any minimum qualification, is mistakenly classified as a non-master, or in which someone who is wholly incompetent is mistakenly classed as a master.

Where attention to cutoffs must be given, the introduction to the topic by Hambleton et al. (1978) is worth noting:

> "The problem of determining cutoff scores for assigning examinees to mastery states based on their criterion-referenced test performance has received much attention from researchers in recent years. Still, the problem seems far from resolved. The

arbitrariness of the proposed solutions has proved troubling to some measurement people, to the point where they seriously question the merits of determining and using cutoff scores at all (p. 26-27).

Despite the psychometric futility of cutting scores, there is no administrative procedure more widely accepted for test score interpretation than the establishment of some sort of cutoff score. Moreover, for some procedures of evaluating the usefulness of a test, standard reference points on the score distribution, similar to cutting scores in the way they are chosen, are necessary. The problem will not go away just because it is psychometrically intractable.

Four different ways can be used to establish such scores. All involve judgments, and each involves the collection of some kind of data as a basis for the judgment.

1. Normative determination. Although job knowledge and other work sample tests are generally intended to lead to content-referenced interpretations, cutoff scores may nevertheless be norm-referenced interpretations. One may determine a priori that some specified percentage of a specified sample of people can be assumed to be masters. These percentages are typically exercises in imagination, drawn from the air. Once the judgment has been made, however, the determination of a cutting score requires a distribution of scores from an appropriate sample of people.

2. Absolute decisions. It is sometimes argued that, if the job content domain has been properly sampled, and if the test content domain has added no irrelevancies, anyone who is really a master of the content domain should be able to pass all items in the tests. This means that the cutting score for mastery is set at the absolute value of 100%. Failing even one item results in being classified as a nonmaster. The arrogance in establishing such a cutting score is clear when one considers that even the most carefully devised examination is subject to some error of measurement. Whether one computes a standard error of measurement or a

standard error of estimated ability, or whether one simply arbitrarily makes an allowance for error, some departure from the 100% mark is allowable as a permissible degree of error. Even in these circumstances, however, the 100% score is the intended mastery score; the actual cutoff score is placed at something less than 100% to take into account computed or judged errors of measurement.

3. Decision theory. Decision theoretical models have been established for cutting scores. These can be illustrated with reference to Figure 3. If it can be assumed that Type 1 errors of classification are equal in importance and cost to Type 2 errors of classification, an optimal cutting score -- that is, one which minimizes the errors of classification -- is the point at which the distributions of scores of masters and the distribution of scores of non-masters intercept. Rorer, Hoffman, Laforge, and Hsieh (1966) and Rorer, Hoffman, and Hsieh (1966) have provided procedures for determining these cutoff scores with other loss functions.

4. Estimation of item difficulties. Ebel (1972) has proposed a knowledge estimation procedure for setting cutting scores. The procedure requires that the items be classified both in terms of importance level and difficulty level. Within each group of items so defined, knowledgeable judges determine or estimate the number of minimally qualified people who will get the items correct. The passing score is a percentage based on the average of these estimates. Essentially this same procedure was used by the Educational Testing Service in determining cutoff scores for certification of teachers in the State of South Caroline (Educational Testing Service, 1976); the legal success of the procedure was established by the fact that this usage was accepted by the courts up through the Supreme Court.

It should be emphasized again that none of these methods has special merit particularly from a measurement perspective. What is required is not a statistically or psychometrically defensible method, but rather a consensus among knowledgeable judges that a specific procedure, and the result of that procedure, is justified.

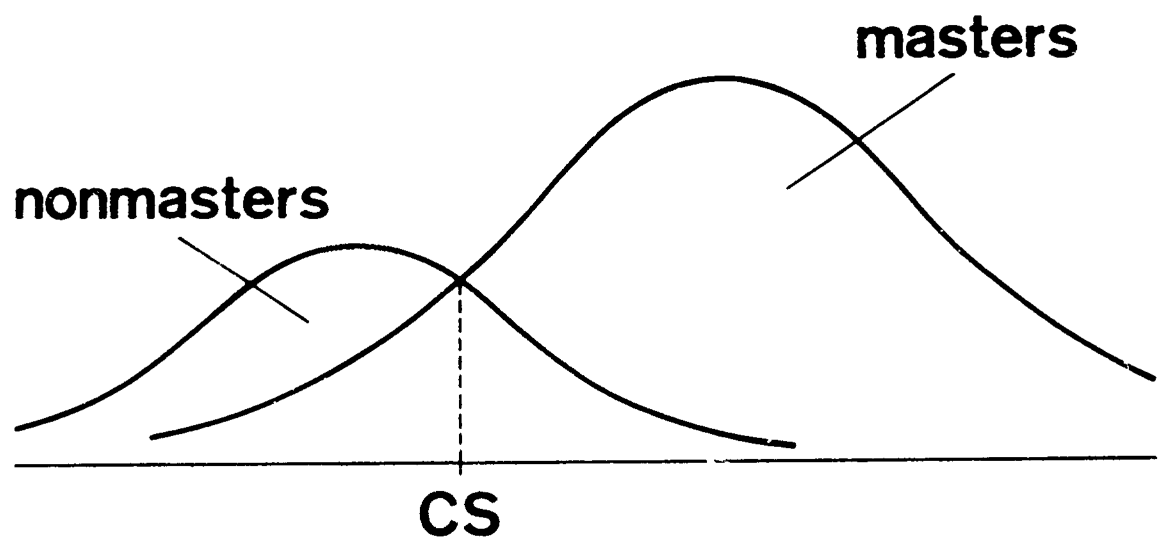

Figure 3. Cutting score or point of intersection of score distribution of masters and of nonmasters

GENERALIZABILITY ANALYSIS

Generalizability theory, as described by Cronbach et al. (1972), is an application of the principles of analysis of variance to estimate the magnitudes of specified sources of error in measurement. The simplest design for a generalizability study is a person-by-item design. That is, variance estimates are obtained for individual differences across people and for individual differences across items. An interaction term is also possible, but this is confounded with random error and becomes an estimate of random error variance. Much of the total variance in a set of test scores should be attributable to individual differences among people, but some of it is due to differences in samples of items. The generalizability analysis in this case identifies total variance as a proportion of variance due to individual differences among people, variance due to a main effect of items, and variance due to error.

More complex analyses can be designed for specific situations. In one common kind of situation, a group of people is given a work sample test in each of several different installations, in each of which there may be several different observers. In analysis of variance terms, persons are nested within observers who in turn are nested within installations. Complicating the situation is the frequent assumption that work sample observations made under conditions of probationary judgments generalize to a later time when the work sample is observed under less severe institutional control. A check on the assumption requires two conditions of measurement. The necessary research design can be described in Figure 4 with a Venn diagram identifying the potential sources of variance, overlapping circles indicating interactions. Using the notation of Brennan (1977), the design can be identified as p:o:i x c, persons nested

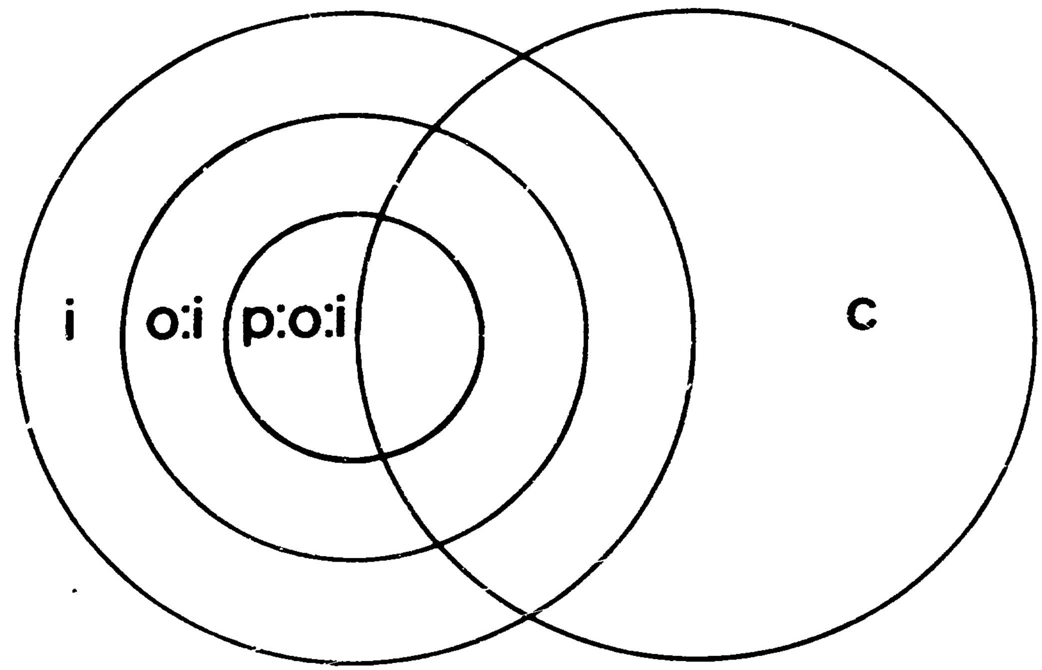

Figure 4. Venn diagram identifying variance estimates for persons (p) nested in observers (o) nested in installations, all crossed with conditions. c, i, o:i, p:o:i, i x c, o:i x c, and p:o:i x c

within observers nested within installations by condition. This experimental design makes possible seven estimates of variance sources, one for the total confounding of persons, observers, and installations, another for the confounding of observers and installations, another for the effect of installations alone. It will be recognized, at this point, that no main effect can be determined for persons or for observers independently of installations; however, a main effect for installations is possible; if the main effect for installations is essentially the same as the main effect for the confounding of observers and installations, it may be assumed that differences in installations contribute very little to the overall variance -- which is as it should be. However, if the main effect due to installations is not substantially different from the main effect of the confounding of persons, observers, and installations, then it would also follow that persons are representing very little contribution to total variance --- which is <u>not</u> the way it should be in attempting to measure individual differences.

Continuing the list of contributions to variance, the study will investigate a main effect due to condition and three interaction terms, one for each of the three levels of confounding of persons, observers, and installations by condition. The interaction terms, in each case, may be taken as error terms and, when one is interested in measuring the proficiency of persons, effects due to condition, installation, or observer-installation confounds are all sources of systematic error variance.

The principal usefulness for work sample testing of the paradigm is that it identifies the sources of error and, in so doing, identifies the limits to which scores on a test may be generalized. Suppose, for example, that there is indeed a strong main effect due either to installations or to conditions. It would follow then that scores obtained in one installation under one condition would not generalize to other conditions. It would be possible, using the Cronbach, et al. equations, to develop an estimated observed score for specific combinations of installations and conditions. Insofar as the interaction terms are relatively small, therefore, a test may be said to be useful even across circumstances which have in fact a substantial main effect. If, however, the interaction terms are substantial, the error corrections are of questionable reliability and the test lacks dependability in measurement.

A generalizability analysis could be used to check on such effects as racial bias by using ethnic identification as one of the facets or conditions of measurement. In military applications, a work sample test may be given under highly standardized conditions that are somewhat aseptic. This is desirable from the point of view of precision in measurement, but it poses severe problems from the point of view of the generalizability of the results of that measurement. Does it necessarily follow that people who do well under the aseptic conditions of, for example, a training post will also do as well under the more realistic conditions of performing the same tasks in the field, in the rain, under conditions of jungle heat or arctic cold, or perhaps under combat conditions? While it is highly unlikely that an experimental study will include actual combat conditions, it is logical to design a generalizability study to facet a series of conditions ranging from highly facilitating to highly hostile. The REALTRAIN program (Shriver et al., 1975) offers an example of conditions which are not aseptic but which represent a closer approximation to combat than would be true in, let us say, a training center. Taking the same kinds of observations in both kinds of situations produces an opportunity to determine the degree to which variations in the hostility of conditions account for variance in the overall performance of individuals.

The topic of generalizability is of such great importance that it is singled out for the fourth report of this series.

EVALUATIONS OF WORK SAMPLES

JOB RELATEDNESS

The job relevance of a work sample test, particularly if it is a direct work sample, is almost never seriously questioned. The relevance of a job knowledge test, or even of the most highly abstracted work sample, is usually assured if efforts to assure have permeated all phases of test development from job analysis through the choice of scoring procedures. This is the singular beauty of work sample testing. Of the six requirements for a test to be accepted as an operational definition of a variable (Guion, 1977, in press), only two points of question seem likely to arise.

The first of these is the basic question of the suitability of the test content universe and domain relative to the job content domain. If the judgment is that knowledge or abstractable skill is to be measured, especially if measured by paper and pencil rather than hands-on methods, the test loses its face validity. It may be challenged as being designed more for the convenience of the tester than to test empirically justifiable prerequisites to effective performance. Claims of job relevance are more secure if judgments at this step are considered and explained and the process, reasoning, and degree of consensus are documented extremely carefully.

The other point of possible contention is in the scoring. Scoring procedures for a typical job knowledge test, whether traditional or utilizing latent trait analysis, are likely to be quite straightforward and, of themselves, be subject to little challenge. Such scores may, however, be contaminated by irrelevant variables such as reading ability or specific forms of test wiseness. Plausible hypotheses of contaminants in direct work sample may change the relevance of these tests, too. To the extent that such questions seem reasonable, studies evaluating the construct validity of the test or of the test inferences from the scores may be necessary.

RELIABILITY

Conventional reliability estimates are as applicable to work sample tests as to others. One may compute coefficients of stability or of internal consistency -- remembering, in the latter case, that the internal consistency of content-referenced tests need not be particularly high so long as there is some. Since the recommended procedures have been to include a large enough item pool to permit the generation of roughly parallel forms, even something similar to the classical coefficient of equivalence can be computed although the correlation of independently constructed domain samples is not a conventional estimate of reliability since they are not strictly parallel.

Reliability of Mastery Classification. There are also some special problems in estimating reliability for content-referenced tests. Many special methods have been proposed; many of them really yield estimates of the reliability of mastery classification schemes rather than estimates of the reliability of measurement (Livingston, 1976). The kappa statistic for correlating nominal data has been used, with modifications, by Huynh (1976) and by Swaminathan, Hambleton, & Algina (1974) for determining an estimate of percentage of agreement in classification that takes into account the agreement that might be expected solely by chance.

It should be recognized, of course, that the reliability of classification depends on the placement of the cutting score; Huynh has pointed out a non-linear relationship between kappa and the cutting score. As the cutoff score gets larger, kappa increases up to some maximum, after which it decreases. Test length and test variability also influence the kappa statistic. Subkoviak (1976) presented a different method of estimating the reliability of classifications, a coefficient of agreement defined as the probability that an individual will be assigned to the same mastery state on parallel tests. The definition does not require a limitation of only two mastery states.

Reliability of Measures. The reliability of the scores as measurement, rather than as a basis for decision, has been approached by different authors in somewhat different ways, but commonly involving discrepancies between the obtained score and the point on the score distribution defined as the standard or cutting score. The analogy to norm-referenced testing is straightforward; in conventional classical reliability estimation, one is basing the estimate on the discrepancies between obtained scores of individuals and the mean of the distribution. Many of these techniques are essentially equivalent to classical reliability if the cutting score happens to be at the mean.

Livingston (1972) developed an equation for computing the reliability of measurement using the equation

$$r_{cc} = \frac{r_{xx}S_x^2 + (\bar{X}-C)^2}{S_x^2 + (\bar{X}-C)^2}$$

where

r_{cc} is the content-referenced correction of r_{xx},

r_{xx} is the classical reliability coefficient,

S_x^2 is the classical test variance,

$\bar{X}$ is the test mean, and

C is the standard or cutting score.

More recently, he has extended the basic idea to provide a statement of the reliability of a single score (Livingston, 1976). This modifies the above equation by setting r_{xx} at zero (to characterize a joint probability distribution of variation due to irrelevant conditions) and by replacing the mean of a distribution with an estimate of the individual's true score. The resulting equation is

$$r_{ss} = \frac{(T-C)^2}{S_e^2 + (T-C)^2}$$

where

r_{ss} is the reliability of the single score of one examinee,

S_e^2 is the error variance, the square of the standard error of measurement,

T is the examinee's estimated true score, and

C is the standard or cutting score.

The reliability estimate in this case is, therefore, somewhat analogous to a statistical test of significance in that the reliability depends on how far an individual's true score is from the criterion score. If the true score and criterion score are identical, the reliability in the single score is given as zero.

The estimate of the reliability of content-referenced measures recommended by the present author has been developed by Brennan and Kane (1977) and combines the Livingston approach and generalizability theory. Brennan and Kane pointed out that the $\bar{X}$-C term is fundamentally concerned only with errors due to sampling people; if one is also concerned with errors due to sampling items, as implied by the entire notion of domain sampling, a somewhat more complex determination is needed. To be consistent with generalizability theory, they prefer to call the measure they have proposed an <u>index of dependability</u> rather than an estimate of reliability.

The classical assumption in psychometric theory is that an obtained score equals a true score plus an error score. Brennan and Kane start from a more complex linear model:

$$X_{pi} = \mu + \pi_p + \beta_i + \pi\beta_{pi} + e_{o(pi)}$$

where

μ is the grand mean in the population of persons and universe of items,

π_p is the effect for person p,

β_i is the effect for item i,

$\pi\beta_{pi}$ is the effect for the interaction of person p and item i, and

$e_{o(pi)}$ is the error with o representing a replication subscript

Since the ordinary case of affairs provides only one observation for each person-item combination, the error term and the interaction term are totally confounded and can for practical purposes be combined.

From this initial assumption that an obtained score is a deviation from a grand mean of possible scores according to the influence of the characteristics of the person being measured and the characteristics of the items chosen for measuring him, plus the ubiquitous error, Brennan and Kane derive an equation for estimating the index of dependability, noted here as I_d.

If the items are scored along a continuous scale, an analysis of variance procedure can be used to estimate I_d. Where the items are scored dichotomously, and the estimation can be simply expressed as

$$I_d = 1 - \frac{1}{n_i - 1}\left[\frac{\overline{X}_{PI}(1 - \overline{X}_{PI}) - S^2_{pi}}{(\overline{X}_{PI} - C)^2 + S^2_{pi}}\right]$$

where

I_d is the index of dependability,

n_i is the number of items,

$\overline{X}_{PI}$ is the grand mean of the item scores (which are either 0 or 1) over all persons and all items,

C is the standard or cutting score, and

S^2_{pi} is the variance of the mean scores of persons over items.

The essential and important feature of this approach is its consideration of errors of sampling people and of sampling items. For content-referenced measurement, this double consideration is very important. The final test administered to an individual consists of only a sample of the items that might have been developed from a domain, a fact with special importance when scores are to be interpreted with reference to that domain.

A Coefficient of Accuracy. Before leaving the topic of reliability, an interesting new statistic specifically designed for content-referenced testing is the coefficient of accuracy proposed by Shaycoft (1977). It is not a reliability coefficient but is an analog of reliability. Using the example of a clinical thermometer with a systematic 2° error, Shaycoft argued that the thermometer would not be satisfactory for its purpose even if the readings it yielded were perfectly reliable. By analogy, a content-referenced test containing systematic error in the scores might yield a high reliability coefficient because of the systematic error. These observations suggest a need for a new psychometric statement which she called a coefficient of accuracy, analogous to a reliability coefficient except that it is reduced by any measurement error, be it random or systematic.

She also proposed an accuracy analog for the standard error of measurement, and provided a basis for correcting the coefficient of accuracy for variations in range. With no experience using the coefficient of accuracy, the writer is unprepared to make recommendations about it other than to suggest that its implications be studied; the concept has potentially great practical significance for content-referenced testing.

Modern Replacements for Reliability. Particularly in direct work sample tests, the errors due to differences among observers may account for a substantial portion of the error variance in overall scores. Moreover, work sample testing is usually conducted primarily for the purpose of identifying maximum potentiality; for this reason, some variance might be due to differences in motivational arousal, such as differences between conditions of institutional control, which tend to maximize motivation to perform, and conditions of field observation, in which the motivation to perform may be not quite so high.

Neither kind of differences, however, fit under the usual rubric of reliability estimation; they are much more appropriately described under the generalizability rubric.

The generalizability of measurement is an essential characteristic in work sample testing. Work sample observations which are highly dependent upon the place, time, or condition of observation, or on who does the observing, are less useful than those that can be depended upon to give common results under different conditions. It is argued, therefore, that generalizability, as the more general case in reliability estimation, is the critical consideration in the evaluation of work sample testing.

Latent trait analysis produces a precision estimate that also supersedes conventional reliability, the information function. Latent trait analysis does not preclude the use of a generalizability study. The primary value of latent trait analysis is as a scaling procedure; generalizability analysis investigates the sources of error in obtained scores, regardless of the metric used for describing those scores. Generalizability theory estimates the magnitude of various errors; latent trait theory minimizes error.

A content-referenced test, constructed and scored according to the principles of latent trait analysis, can be shown from the test information curve to have specific limits of probability error at different score levels. This is an estimate of the degree of precision in measurement; the precision of the estimate itself cannot be matched by any of the conventional techniques of reliability estimation.

VALIDITY

Notions of validity seem almost superfluous in discussions of work samples, at least of "total job" or "direct" work samples. For many such tests, only so-called content validity, described in these reports as job relatedness, seems important.

The evaluation ordinarily meant by the term content validity has already been described here as job relevance. The essential evidence of the validity of scores on a content sample stems not from statistical analysis of the scores themselves but from an evaluative analysis of the judgmental processes involved in developing the test. As Messick (1975) has pointed out, what has been called content validity is better described as content-oriented test development.

The panels of judges who evaluate items in the item analysis are at the same time providing the fuel for the evaluations of the test as a whole. The job relatedness ratio for items, which Lawshe (1975) called a content validity ratio, can be turned into a job relatedness index, which is simply the mean of the job relatedness ratio values of the items in the final form. It is an index number, expressed on a scale from -1 to +1, describing the degree to which knowledgeable expert judges perceive overlap between the ability to do a job and the ability to answer the items of the test correctly.

Construct Validity. Part of what people mean by content validity is a special case of construct validity. Therefore, disconfirmatory studies need to be conducted to investigate the possibility that variance in a set of test scores is attributable to characteristics other than the knowledge or proficiency inferred from the scores. For a job knowledge test, this is probably not a terribly serious problem. However, as pointed out in the section on job relatedness, one might encounter the criticism of a given job knowledge test that scores on the test are as influenced by ability to read as by knowledge of the job. If the job content domain includes a great deal of reading, such a criticism is not likely to be seriously proposed. If, however, the job is one in which little or no reading is involved, then a job knowledge test consisting of multiple-choice items is likely to be considered contaminated by including a component not in the job content domain and therefore an irrelevant source of variance. The logic of construct validity needs to be invoked in evaluating this possible interpretation of obtained scores.

It might be done in either of two ways. One might determine the readability level of the test and decide whether it is high enough to reduce the scores of people who are otherwise qualified to take the job. Or, scores on the test can be correlated with scores on a standard measure of reading proficiency If the correlation is high, the alternative interpretation of scores on the job knowledge test is supported. If it is low, however, it indicates that only a small proportion of the total variance in the test scores can be attributed to reading ability.

A common problem in content-referenced measurement is that an available sample is very likely to have low variance among scores; construct validity studies based on correlational research will yield low correlations in such samples, failing to show relationships that may actually exist. For this reason, among others, it has been argued that content-referenced tests should be developed to yield a reasonable spread of scores.

Consistency in Domain Sampling. In discussing what was called content validity, Cronbach (1971) suggested the independent construction of two tests from the same content domain and test specifications. The recommendations for test development in this report called for developing substantially more than twice as many items as might be needed. This makes it possible to conduct some item analysis research, identify and discard poor items, and still have a sufficiently large item pool to permit the allocation of items into two forms. (This is not, of course, precisely the operation prescribed by Cronbach, but it does provide similar samples of the domain.) The validity of the sampling procedure can be assessed by correlating the total scores on these two independent sets of items. More important is that the forms provide an important facet in generalizability studies.

Predictive Utility. While there is no great objection to doing predictive validity studies to determine whether the job knowledge test does in fact predict future performance, neither is there any particularly good reason for doing so. While a high predictive validity coefficient between scores on a work sample and some later measure of performance suggests additional evidence of the construct validity of inferring proficiency from work sample scores, a low predictive validity coefficient would not cast doubt on the validity of such inferences. Rather, it would cast doubt on the validity of inferences from the criterion measure; unless the criterion is another work sample, or the same work sample observed under other conditions (in which case we are discussing a generalizability model), the criterion is unlikely to be as carefully constructed or as job related as the work sample itself. It is true that the logical foundation for using a job knowledge test for placement decisions is an implied prediction. Nevertheless, the implied prediction is simply that, if a high scoring person is placed on a job, he will be able to do the job immediately.

Criterion-related validation can be useful in evaluating abstracted work samples or job knowledge tests. Although abstractions, such tests are intended to be used for inferring proficiency. Beginning with the definition of a job content universe to the definition of a test content domain, the assumption and development of a job knowledge test or other abstraction has consistently been that the knowledge or skill being tested is an essential prerequisite to successful performance on the job. This is a hypothesis; it is a hypothesis that performance on one variable, such as job knowledge, is related to performance on a different variable, proficiency. This is a correlational hypothesis, and it should be tested following the principles of criterion-related validation. Such a study may not be necessary if there is sufficient agreement among qualified judges that the knowledge tested is indeed prerequisite to effective performance.

SUMMARY: PRINCIPLES OF WORK SAMPLE TESTING

This paper has concentrated on the application of general principles of measurement theory and on general principles of the evaluation of psychological measurement to the construction and evaluation of work sample measures of performance proficiency. It started with an emphasis on job analysis as the foundation for work sample test development, and it continued with idealized suggestions for test construction and evaluation for various degrees of work sample abstraction, that studies may be needed to demonstrate that performance on an abstract work sample is consistent with performance on the job itself. The greater the degree of abstraction, the greater the necessity for empirical verification of that relationship.

In summary, seven principles can be drawn from this report which should be observed i . the evaluation of work sample tests:

1. The choice of a job content domain needs to be justified. The job content domain is a portion of a total job content universe. It is chosen on the basis of expert judgment, not because it is representative of the universe, but because it is an important or salient aspect of the universe for the purposes of the decisions that are to be based on the testing. In a situation in which some aspects of the job content must be learned after one is placed on the job (Gael, 1977), a selection or certification test should not include those portions of the job content universe. Rather, it should be restricted to a domain of job skills or knowledge or activities that a candidate for the job is expected to bring to it. Other kinds of purposes may impose other kinds of restrictions. For skill qualification testing, the restrictions should probably be minimal; that is, to certify competence to perform a job, the job content domain should probably be close to the total job content universe. It might still, however, represent a substantial amount of abstraction from that universe to eliminate redundancies, or to emphasize the most important aspects of the universe, or for other reasons. Any condensation of the job content universe into a smaller job content domain should be accompanied by careful documentation of the reasoning used, and it should also record the reliability of the independent judgments of panel members and the degree of consensus achieved through panel discussions.

2. The test content domain and the job content domain should be as congruent as possible. Measurement of that congruence is a matter of judgment. Numerical indices like Cronbach's correlation of independently developed content samples, the Lawshe content validity index, or the Rovinelli-Hambleton index of test-objective congruence may all be used, but none of them should be taken too seriously. For one thing, the two domains are defined in abstract terms and any actual attempt to make them literally measurable would probably distort them. For the other, the important evidence consists of the judgments of qualified experts, not of statistical indices. Such indices are valuable, but they are valuable precisely because they provide a means by which expert judgments may be systematically collected and analyzed.

3. Scoring procedures should approximate formal, fundamental measurement as much as possible. Special care should be taken to assure that scores are indeed established along a transitive scale of measurement. In practice, any set of

numbers ultimately assigned is, of course, mathematically transitive; that is, if individual a gets 10 points, b gets 15 points, and c gets 20 points, then obviously b has a higher score than a, c has a higher score than b and also has a higher score than a. If, however, the 10, the 15, or 20 points earned are earned on quite different bases, it does not follow that b is necessarily more proficient than a, or that c is more proficient than b, or that c is more proficient than a, even if the other two statements were true. Obviously, the call for transitivity is a call for reasonable homogeneity or functional unity in the system of scoring. If this unity cannot be achieved, individual subsets of scores should be obtained; the importance of having such subsets can be inferred from their intercorrelations. It is better to try to develop the subsets and find out they are not necessary because of high intercorrelation than simply to assume that the overall score is sufficiently homogeneous.

4. Levels of proficiency should be measured: scores should not be merely dichotomies. Test construction, particularly following the recommendations of this report, is too expensive in time and resources to allow the deliberate loss of information caused by dividing an entire distribution of scores into just two parts. By using continuous scoring, a test can continue to be used even when standards change (as for different levels of a common career ladder or in response to new circumstances). Moreover, most of the appropriate procedures for evaluating the test -- reliabilities, information functions, construct validities, generalizability coefficients -- require variance along a continuous scale of measurement.

5. The opportunities for irrelevant influences on individual scores should be at a minimum. Insofar as all testing procedures are carefully standardized, and insofar as there is some variability in performance, this principle is essentially an admonition to check for violations of the assumption of construct validity. It is better, however, to recognize that work sample tests may, in a well prepared sample of subjects, result in very low variances, in which case investigations into alternative construct explanations are difficult. Even in these cases, however, it is very important to make the effort to ascertain the possible influence of attributes of observers, attributes of conditions, or simply irrelevant attributes of the people being measured on the performance of the work sample.

6. Work sample scoring, if the test is to be used in an organization with diverse locations, should be standardized on a scale of reference that is applicable to an organization-wide population. Some form of content standard score is preferable to a standard normative scale of measurement. While normative interpretations may not be irrelevant for many uses, the principal meaning of a set of work sample scores should be inherent in the content of the test. Number or proportion of items answered correctly, number of points awarded on an observer scoring form, or simply a dichotomous pass-fail notation may be seen as roughly a content standard score. Nevertheless, more sophisticated procedures can and usually should be used, such as latent trait scaling, or keying scores to special Guttman scales, or functional scaling.

7. Scores from work sample testing in the usual conditions of institutional control must generalize not only to field settings but to a variety of field settings. Work sample testing, before it is made operational on a large scale, should be subjected to appropriate generalizability analysis.

The recommendations of this report for work sample testing are extensive; following all of them for very many jobs would be prohibitively expensive and laborious. The use of a panel of experts from the beginnings of job analysis to the completion of test development would come to countless manhours of deliberation. The recommended calibration by latent structure analysis and evaluation by generalizability studies could cost many thousands of dollars for data collection alone. The expenditures of time and money may be justified for certain extremely critical jobs, but not for many. Moreover, the world of measurement does not contain enough test development specialists to conduct such intensive campaigns for effective work sample measurement for more than a few very important job categories.

The conclusion is inescapable that short cuts are needed. If a work sample were developed according to all of the recommendations in this report, its job relevance and (where it matters) its validity would be unquestionable. But will this be true if short cuts are taken? It probably will be with some short cuts, but we don't know which ones are "safe." The report has spoken repeatedly of abstracting from universes and domains; the job of developing work sample tests has a total job content universe as does any other, and short cuts represent similar abstracting. What short cuts or abstractions will yield work samples, direct or abstract, as unquestionable in relevance and validity as those developed without cutting corners?

It is an empirical, not a rhetorical, question. Systematic studies are needed to compare work samples developed by simpler procedures to model work samples developed by the more elaborate procedures outlined in this report.

The most important questions, however, are not questions of procedural simplification. They are questions of the optimal or permissible kinds and levels of abstraction from a job or test content domain in the development of abstract work samples or job knowledge tests. It is a problem in generalizability. Will abstractions of one, developed by one set of rules, generalize more or less well than abstractions of a different kind, or developed by a different set of rules, to model work samples developed with very little abstracting at all?

REFERENCES

Anderson, N. H. How functional measurement can yield validated interval scales of mental qualities. Journal of Applied Psychology, 1976, 61, 677-692.

Asher, J. J., & Sciarrino, J. A. Realistic work sample tests; a review. Personnel Psychology, 1974, 27, 519-533.

Bejar, I. I. An application of the continuous response level model to personality measurement. Applied Psychological Measurement, 1977, 1, 509-521.

Boyd, J. L., Jr., & Shimberg, B. Handbook of performance testing. Princeton, N.J.: Educational Testing Service, 1971.

Brennan, R. L. Generalizability analysis: Principles and procedures. (ATC Technical Bulletin No. 26). Iowa City: American College Testing Program, 1977.

Brennan, R. L., & Kane, M. T. An index of dependability for mastery tests. Journal of Educational Measurement, 1977, 14, 277-289.

Brenner, M. H. An experimental comparison of cluster-analysis-refined spatial relations items constructed by two methods: Gutman scale analysis and item analysis. Unpublished masters thesis, Bowling Green State University, 1959.

Campbell, J. T., Crooks, L. A., Mahoney, M. H., & Rock, D. A. An investigation of sources of bias in the prediction of job performance: A six-year study. Project Report PR-73-37. Princeton, N.J.: Educational Testing Service, 1973.

Comer, J. C. Trade competency testing via specimen inspection. Journal of Industrial Teacher Education, 1971, 9, 50-54.

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.) Educational measurement (2nd ed.) Washington, D.C.: American Council on Education, 1971.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements. New York: Wiley, 1972.

Duffy, P. J. Development of a performance appraisal method based on the duty module concept. (Technical Paper 273). Arlington, Va.: U.S. Army Research Institute for the Behavioral and Social Scien[illegible] 1976.

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

Ebel, R. L. Essentials of educational measurement. Englewood Cliffs, N.J.: Prentice-Hall, 1972.

Educational Testing Service. Report on a study of the use of the National Teachers Examinations by the state of South Carolina. Princeton, N.J.: Educational Testing Service, 1976.

Equal Employment Opportunity Commission. Guidelines on employee selection procedures. Federal Register, August 1, 1970, (No. 149), 12333-12336.

Federal Executive Agency Guidelines. Uniform guidelines on employee selection procedures. Federal Register, December 30, 1977, (No. 251), 65542-65552.

Foley, J. P., Jr. Overview of Advanced Systems Division Criterion Research (maintenance). Paper presented at Symposium on Criterion Development for Job Performance Evaluation, San Antonio, Texas, 23-24 June, 1977.

Gael, S. Development of job task inventories and their use in job analysis research. Catalog of Selected Documents in Psychology, 1977, 7(1), 25. (Ms. No. 1445).

Grant, D. L., & Bray, D. W. Validation of employment tests for telephone company installation and repair occupations. Journal of Applied Psychology, 1970, 54, 7-14.

Guilford, J. P. Printed classification tests (Report No. 5). Washington, D.C.: Government Printing Office, 1947.

Guion, R. M. Content validity -- the source of my discontent. Applied Psychological Measurement, 1977, 1, 1-10.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48. 1-47.

Hively, W., II, Patterson, H. L., & Page, S. H. A "universe-defined" system of arithmetic achievement tests. Journal of Educational Measurement, 1968, 5, 275-290.

Huynh, H. On the reliability of decisions in domain-referenced testing. Journal of Educational Measurement, 1976, 13, 253-264.

Ironson, G. H. A comparative study of several methods of assessing item bias. Unpublished doctoral dissertation, University of Wisconsin-Madison, 1977.

Jones, A., & Whittaker, P. Testing industrial skills. New York: Wiley, 1975.

Lawshe, C. H. A quantitative approach to content validity. Personnel Psychology, 1975, 28, 563-575.

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

Livingston, S. A. The criterion-referenced reliability of a single score. (COPA 76-01). Princeton, N.J.: Educational Testing Service, 1976.

Lord, F. M. Individualized testing and item characteristic curve theory. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), Contemporary developments in mathematical psychology, Vol. II, San Francisco: Freeman, 1974.

Lord, F. M. Practical applications of item characteristic curve theory. Journal of Educational Measurement, 1977, 14, 117-138.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Maier, M. H., Young, D. L., & Hirshfield, S. F. Implementing the skill qualification testing program. (R & D Utilization Report 76-1). Arlington, Va.: U.S. Army Research Institute for the Behavioral and Social Sciences, 1976.

Messick, S. The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 1975, 30, 955-966.

Millman, J., & Popham, W. J. The issue of item and test variance for criterion-referenced tests: A clarification. Journal of Educational Measurement, 1974, 11, 137-138.

NCARB. NCARB Architectural Registration Handbook. Washington, D.C.: National Council of Architectural Registration Boards and Architectural Record Books, 1976.

Peterson, D. A., & Wallace, S. R. Validation and revision of a test in use. Journal of Applied Psychology, 1966, 50, 13-17.

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

Root, R. T., Epstein, K. I., Steinheiser, F. H., Hayes, J. F., Wood, S. E., Salzen, R. H., Burgess, G. G., Mirabella, A., Erwin, D. E., & Johnson, E., III. Initial validation of REALTRAIN with Army combat units in Europe. (Research Report 1191). Arlington, Va.: U.S. Army Research Institute for the Behavioral and Social Sciences, 1976.

Rorer, L. G., Hoffman, P. J., & Hsieh, K. Utilities as base rate multipliers in the determination of optimum cutting scores for the discrimination of groups of unequal size and variance. Journal of Applied Psychology, 1966, 50, 364-368.

Rorer, L. G., Hoffman, P. J., LaForge, G. E., & Hsieh, K. Optimum cutting scores to discriminate groups of unequal size and variance. Journal of Applied Psychology, 1966, 50, 153-164.

Rovinelli, R. J., & Hambleton, R. K. On the use of content specialists in the assessment of criterion-referenced test item validity. Tijdschrift voor Onderwijsresearch, 1977, 2, 49-60.

Rubinsky S., & Smith, N. Safety training by accident simulation. Journal of Applied Psychology, 1973, 57, 68-73.

Rudner, L. M. Item and format bias and appropriateness. Washington, D.C.: Model Secondary School for the Deaf, 1976.

Samejima, F. Homogeneous case of the continuous response model. Psychometrika, 1973, 38, 203-219.

Schmidt, F. L., Greenthal, A. L., Berner, J. G., Hunter, J. E., & Seaton, F. W Job sample vs. paper-and-pencil trades and technical tests: Adverse impact and examinee attributes. Personnel Psychology, 1977, 30, 187-197.

Schriver, E. L., Mathers, B. L., Griffin, G. R., Jones, D. R., Word, L. E., Root, R. T., & Hayes, J. F. REALTRAIN: A new method for tactical training of small units. (Technical Report S-4). Arlington, Va. U.S. Army Research Institute for the Behavioral and Social Sciences, 1975.

Shaycoft, M. F. The coefficient of accuracy: A new statistic for criterion-referenced tests. Palo Alto, Calif.: American Institutes for Research, 1977.

Shepard, L. Reporting results of competency-based measurement. Paper presented at meeting of National Council on Measurement in Education, New York, 1977.

Shimberg, B., Esser, B. F., & Kruger, D. H. Occupational licensing: Practices and policies. Washington, D.C.: Public Affairs Press, 1973.

Slinde, J. A., & Linn, R. L. An exploration of the adequacy of the Rasch model for the problem of vertical equating. Journal of Educational Measurement, 1978, 15, 23-35.

Smith, P. C., & Kendall, L. M. Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 1963, 47, 149-155.

Subkoviak, M. J. Estimating reliability from a single administration of a criterion-referenced test. Journal of Educational Measurement, 1976, 13, 265-276.

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

Thurstone, L. L. Primary mental abilities. Psychometric Monograph No. 1, 1938.

Tryon, R. C., & Bailey, D. E. Cluster Analysis. New York: McGraw-Hill, 1970.

Uhlaner, J. E., Drucker, A. J., & Camm, W. B. Army research in the criterion area: An overview. Paper presented at Symposium on Criterion Development for Job Performance Evaluation, San Antonio, Texas, 23-24 June, 1977.

Wesman, A. G. Writing the test item. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D. C.: American Council on Education, 1971.

Wheaton, G. R., Fingerman, P. W., & Boycan, G. G. Development of a model tank gunnery test. (Technical Report TR-78-A24). Alexandria, Va.: U.S. Army Research Institute for the Behavioral and Social Sciences, 1978.

Wood, R. L., & Lord, F. M. A user's guide to LOGIST. Princeton, N.J.: Educational Testing Service, 1976.

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

Wright, B. D. Sample-free test calibration and person measurement. In Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, N.J.: 1968.

Wright, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

Wright, B. D., & Mead, R. J. BICAL: Calibrating items and scales with the Rasch model. (Research Memorandum No. 23A). Chicago: Statistical Laboratory, Department of Education, The University of Chicago, 1978.

Wright, B. D., & Stone, M. H. Best test design: A handbook for Rasch measurement. Chicago: University of Chicago and Social Research, Inc., 1978.

Yerkes, R. M. (Ed.), Memoirs of the National Academy of Sciences. Vol. XV, Psychological examining in the United States Army. Washington, D.C.: Government Printing Office, 1921.