GENERAL DISTRIBUTION

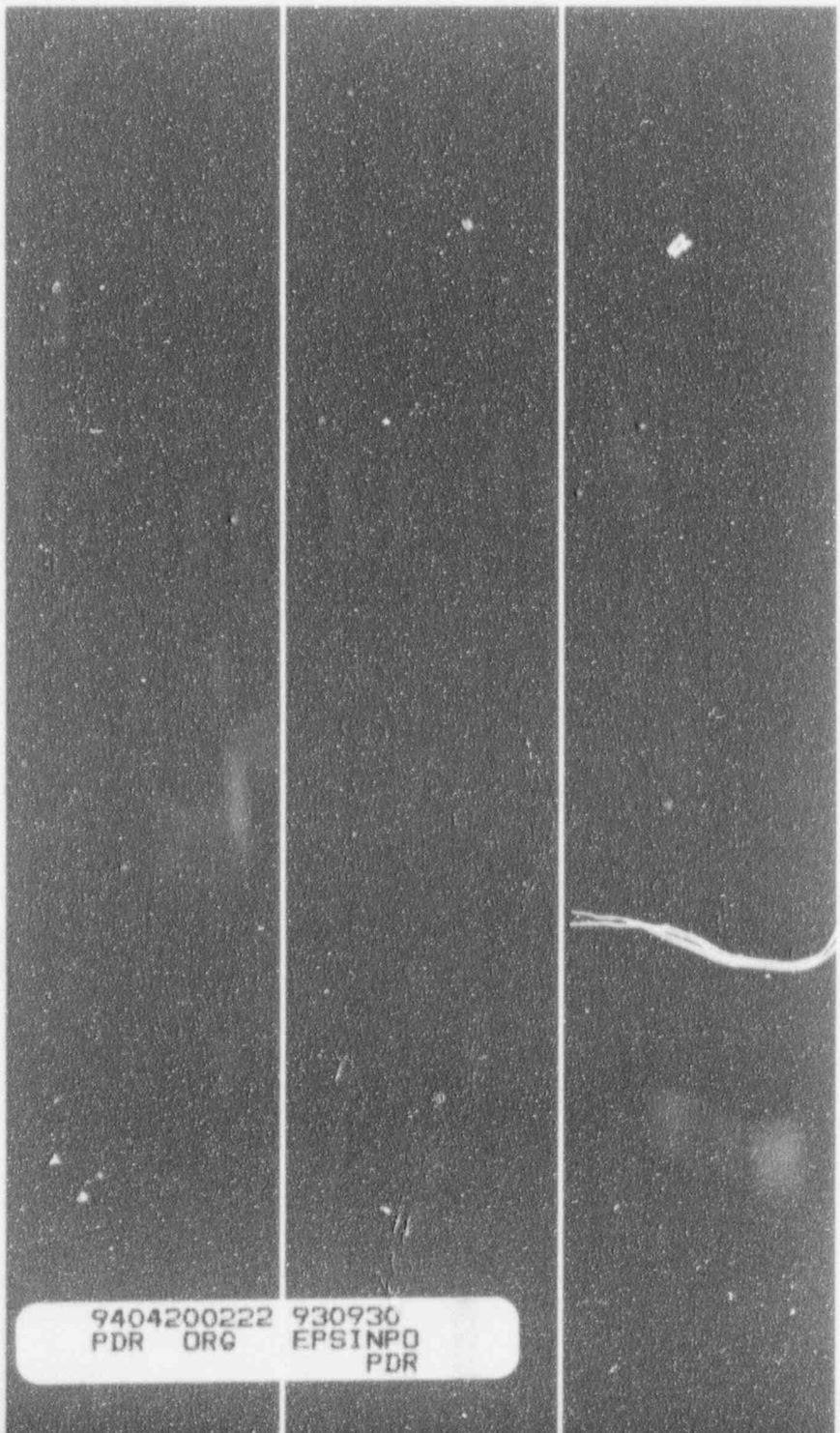
September 1993
ACAD 88-002 (Addendum II)
ACADEMY DOCUMENT

# The Principles of Training System Development Manual, Addendum II: Examinations: Design, Development, and Implementation

NATIONAL ACADEMY FOR NUCLEAR TRAINING

GENERAL DISTRIBUTION

September 1993
ACAD 88-002 (Addendum II)
ACADEMY DOCUMENT



# The Principles of Training System Development Manual, Addendum II: Examinations: Design, Development, and Implementation

NATIONAL ACADEMY FOR NUCLEAR TRAINING

# PRINCIPLES OF TRAINING SYSTEM DEVELOPMENT ADDENDUM II EXAMINATIONS: DESIGN, DEVELOPMENT, AND IMPLEMENTATION

September 1993
ACAD 88-002

NATIONAL ACADEMY FOR NUCLEAR TRAINING

Plant Area: Training and Qualification

Key Words: Examination, Development, Use, Interpretation

The National Academy for Nuclear Training operates under the auspices of the Institute of Nuclear Power Operations (INPO). The Academy provides a framework for a unified, coordinated industry approach to achieving and maintaining effective training and qualification. It also promotes pride and professionalism of nuclear plant personnel. The Academy integrates the training efforts of all U.S. nuclear utilities, the activities of the National Nuclear Accrediting Board, and the training-related activities of INPO.

# TABLE OF CONTENTS

# LIST OF ILLUSTRATIONS

# CHAPTER 1

## A. INTRODUCTION

The development of examinations in the nuclear power industry is both critical and time consuming. Utilities spend a significant amount of their training resources testing trainees. Tests are used for employee selection, qualification, requalification, and promotion. Ineffective testing, or inappropriate interpretation of test results, can have a significantly detrimental effect on personnel performance and plant operations. Test development requires unique skills, and, as with any skill, training and experience are needed to develop proficiency. Like all other aspects of training system development, test development, test use, and test refinement must be part of a systematic process.

The purpose of this addendum is to provide guidance to utility personnel in the development, use, and interpretation of trainee examinations. Addendum I discusses the subject of developing effective test items. This addendum provides guidance in the broader areas of test design, development, implementation, and evaluation. For some test developers this addendum will provide a review of ideas and principles with which they are already familiar; for others it will present new concepts. While this document cannot provide in-depth coverage of test design and development, it should provide the instructor or curriculum developer with a foundation on which to develop sound examinations. This addendum discusses the steps necessary for effective test development. Chapter 2 discusses several uses of tests, and the major types of measuring instruments. Chapter 3 describes the test development process from the initial planning stage, through development and implementation, to scoring and evaluation. Chapter 4 covers several major test and measurement concepts including reliability, validity, and item analysis. It will be a review for the experienced measurement specialist, but for the beginning test developer, many new concepts and terms are presented. It is crucial that the test developer understand and apply these concepts in the development and use of trainee examinations. References are provided to identify sources and provide additional resources for study. Chapter 5 discusses test score interpretation, an essential ingredient in trainee evaluations. Following the text is a glossary of testing terms.

# CHAPTER 2

## A. THE PURPOSES OF TESTING

Testing is normally thought to be an evaluation activity to assess what a trainee has learned. However, this is only one of the purposes of testing in a training program. There are several valid reasons for using tests in job and training environments. Six of these reasons are discussed below.

### 1. Trainee Assessment

By far, the most common use of testing is to determine trainee progress. The instructor combines test scores with other evaluation results to assess an individual's ability to perform a specific job or task. Examples of other evaluation results include the accuracy of trainee responses to oral questioning, the quality of the questions that trainees ask, the competence exhibited in the classroom, laboratory, or work place, and evaluations made by other instructors. While these sources are collected informally, the written test is often the only documented data. Testing, when properly developed and conducted, provides a valid, reliable, and unbiased indicator of trainee performance. Tests, whether written, oral, or performance, provide the most complete and efficient method of collecting data on trainee performance.

What most people commonly think of as testing is really made up of two distinct components: measurement and evaluation. Testing is a **measurement** activity. The purpose of this measurement activity is to score a sample of trainee performance, from which a decision can be made regarding the trainees' performance capability. This decision is the result of an **evaluation** activity. Evaluation is the process of judging the quality of trainee knowledge or skill. This is an important distinction because some individuals incorrectly believe that test **scores** assess performance. Thinking this, they fail to properly interpret the test scores. It is also important to remember that test score interpretation is directly related to the purpose of testing. The two must not be separated.

### 2. Trainee Selection and Placement

Tests are useful for trainee selection and placement. Entrance examinations may be used to waive specific training segments. Test scores may also indicate the need for remedial training. Many utilities use aptitude tests to place trainees in appropriate training programs. Some use interest inventories and psychological evaluations to aid in job placement. Plant management, personnel departments, and training staff must make these decisions based, in part, on test scores.

### 3. Trainee Motivation

Tests are powerful motivators. Trainee study habits are greatly affected by examination schedules. In training programs in which tests are given on a daily or weekly basis, trainees tend to study in anticipation of those tests. Likewise, when there is only an end-of-course exam, many trainees tend to postpone studying until just before the exam. Trainees are also motivated by the feedback a test score provides. Low test scores generally raise trainee anxiety levels, which, if properly channeled, can result in increased concentration and study.

4. **Instructional Improvement**
   Test results can provide rapid feedback regarding instructional effectiveness. If an instructor fails to cover a topic in a classroom presentation or laboratory exercise, lowered test results will normally reveal the ommission. Uniformly high scores for a topic or subject area generally indicate that instruction has been well presented and was effective. Conversely, low scores may indicate that improvement in instruction, teaching material, or strategy is needed.

5. **Program Evaluation**
   When trainee test scores are combined, the data becomes group data. From group data, course or program performance information can be obtained. This information is valuable in assessing program strengths and weaknesses. To maximize the effectiveness of the test data, periodic, systematic reviews should be conducted. This is best done by reviewing item analysis results from objective tests, essay test results, and results from supervisor and trainee feedback questionnaires. With the use of these sources, a composite picture of program strengths and weaknesses can be formed.

6. **Testing as Teaching**
   Instructors who view testing as only an evaluation tool often overlook the opportunity to use testing as a learning tool. An application of testing as teaching can be seen in the on-the-job training qualification check-off process. In this activity the trainee is asked to perform a task under the watchful eye of a master performer. If the trainee performs properly, the performance is acknowledged; if not, the trainee is given immediate feedback on what errors were made and the proper steps needed to correct them. This is effective training. Testing can also be effective instruction in a classroom, especially when test results are reviewed with the trainees. An open discussion of an incorrect answer and why that answer was selected can be very educational for both trainee and instructor.

   Feedback and repetition are two instructional techniques that promote learning. Feedback is an essential element in any learning process. The simple act of steering an automobile is impossible without visual feedback to provide steering corrections. Learning too, is impossible without feedback. Each trainee must have accurate and frequent feedback to confirm that learning is occurring. Trainees continually receive feedback during training from peers and the instructor, and by comparing their work to examples. Exam results provide an additional and valuable feedback source. Trainees should be given the opportunity to review the exam, verifying items answered correctly and identifying their errors. All examinations should be reviewed. End-of-course, or final, examinations are probably the least utilized for providing feedback to trainees. They are often given the last day of class, not returned to the trainee, and not reviewed. A more valuable technique would be to conduct the final exam the day before the last day, score them, and spend the last day in review and summary activities.

   Many instructors prefer not to review exams in class because trainees often argue over their answers. Rather than take this negative view, the instructor should capitalize on a trainee's attempt to raise his score. This attempt to raise a score creates an excellent learning opportunity that can be used to enhance trainee knowledge. This can be done by allowing trainees to defend their answers, justifying why the answers should be accepted as correct.

## B. TYPES OF MEASURING INSTRUMENTS

The following types of tests are widely recognized: achievement, general mental ability, aptitude, interest, personality, and attitude. An overview of each of these measuring instruments follows.

### 1. Achievement Tests

Achievement tests are the most common type of test found in the training environment. General achievement tests are routinely developed and used in performance assessment. Most of these examinations are locally developed, and they are used to qualify individuals. Trainees are passed or failed, based on a pre-established cutoff score. Most tests of this type do not identify unique strengths or weaknesses of the trainee.

Tests that point out strengths or weaknesses are referred to as diagnostic achievement tests. While any test can be used diagnostically, those designed specifically for that purpose have questions that assess specific objectives or selected aspects of the training. A diagnostic achievement test can tell which objectives have not been met. Remedial training can then be provided prior to course completion, certification, or licensing.

### 2. General Mental Ability Tests

Tests that measure general mental ability, often referred to as IQ tests, have little application in most training program settings. These tests are not locally developed but are published standardized instruments, usually requiring certification of the person administering the examination. These tests can be one of two types: individual or group administered. They may be oral, written, or performance tests. IQ scores have sometimes been used to set performance goals. Expections, however, often go unmet since individual differences and motivational factors play a significant role in determining performance.

### 3. Aptitude Tests

Aptitude tests are used to predict a person's ability or skill in a specific field or for a given trait. Aptitude tests do this by testing a person's current skill or knowledge and using this as a predictor of future performance. Aptitude tests are useful when counseling requires identification of a "best" placement. Also, when a person has not been successful in a training program or job classification, an aptitude test may help identify a new field of study or endeavor.

Many utilities use aptitude tests for selecting candidates for entry into training programs. Because of the time and expense of training a proficient worker, it makes sense to select the trainee with the highest probability of success. The Plant Operator Selection System (POSS) is an example of a commonly used aptitude test. Another test used by utilities is the Power Plant Maintenance Positions Selection System. These tests have prov. ı useful in the selection of candidates for training programs.

### 4. Interest Inventories

Knowledge about trainees' interests, personalities, and attitudes is helpful in understanding their needs and matching their goals with those of the employer. Interest inventories can be used in conjunction with aptitude tests to make career guidance decisions.

5. **Personality Inventories**
   Personality inventories are instruments for the measurement of emotion, motivation, and attitude. Most utilities use personality inventories to screen potential employees for personality abnormalities (aberrant behavior). A common instrument used for this purpose is the Minnesota Multiphasic Personality Inventory. Persons who do not "pass" such tests may be considered bad employment risks due to abnormal personality traits. Because of security and safety requirements, they would be prevented from having unescorted access to a nuclear plant.

   Personality inventories have other uses. Personality inventories can be used to identify personality traits, and that information can be used to improve interpersonal communication skills among employees. This could be used in team training for control room operators or similar settings requiring group interaction.

6. **Attitude Inventories**
   Attitude inventories are similar to interest and personality inventories, but because attitudes are personal, there is a high degree of value judgment involved in their measurement. Attitude, like free speech, is often considered the prerogative of the employee. Because of sensitivity in this area, these inventories are seldom used in the employment setting. Attitudes are certainly important; e.g., a positive attitude toward safety is essential in a nuclear power plant, but they are generally assessed informally.

# CHAPTER 3

## A. BASES OF THE TEST

All examinations must be based on training objectives. Effective testing requires that learning objectives are carefully selected and classified prior to test development. A model for the classification of educational objectives is useful.

### 1. Taxonomy of Educational Objectives

Several taxonomies have been devised to classify educational objectives. Foremost among these is Benjamin S. Bloom's Cognitive Taxonomy of Educational Objectives.[1] Krathwohl[2] presented a taxonomy for the affective domain, and Simpson[3] has published the best known taxonomy for the psychomotor domain.

Through these taxonomies, the test developer recognizes that various levels of objectives exist and that test items must match the level of the learning objective. Too often instructors test only lower-level objectives when task performance actually requires a significant amount of higher level performance.

A useful taxonomy for categorizing cognitive learning objectives and their accompanying test items follows. This taxonomy includes six levels and is adapted from Bloom's original work.

*Knowledge:* The recall of previously learned material--The following action verbs are commonly used at this level: define, label, list, name, or state.

*Comprehension:* The ability to grasp the meaning of material--This may be shown by translating material, interpreting material, or estimating future events. Obtaining information from charts, graphs, indicators, and procedures illustrates performance at this level. Frequently used verbs include the following: identify, locate, log, obtain, explain, or estimate.

*Application:* The ability to use learned information to solve routine problems--The following verbs are often used at this level: apply, calculate, derive, sketch, manipulate, operate, or solve.

*Analysis:* The ability to break down information into its component parts so that its structure may be understood--It is the first of the "higher order" cognitive objectives. Frequently used verbs include the following: analyze, interpret, classify, diagnose, or troubleshoot.

---

[1] Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R., Taxonomy of Educational Objectives, Handbook I: The Cognitive Domain (New York: David McKay Co., 1956).

[2] Krathwohl, D. R., Bloom, B. S., & Masia, B. B., Taxonomy of Educational Objectives, Handbook II: The Affective Domain (New York: David McKay Co., 1964).

[3] Simpson, E. J., "The Classification of Educational Objectives: Psychomotor Domain," *Illinois Teacher of Home Economics*, 10(4), pages 110-144, 1966.

*Synthesis:* The ability to put parts together to form a new whole--Information is combined to create a solution, plan, or procedure. Example verbs include the following: construct, create, develop, plan, or write.

*Evaluation:* The highest level of cognitive activity--the ability to judge the value of material for a given purpose--It includes the capability to respond effectively to unique conditions or an uncertain environment. The following verbs illustrate this level: critique, defend, evaluate, judge, or predict.

The levels are ordered from simple to more complex. Knowledge, comprehension, and application are the "lower order cognitive" levels and define the mental capabilities most often performed under normal, routine conditions. Analysis, synthesis, and evaluation are the "higher order cognitive" levels. These skills are essential to the performance of troubleshooting tasks and tasks performed under abnormal or emergency conditions.

## B. PLANNING THE TEST

During test planning, decisions must be made regarding the learning objectives to be tested, the amount of emphasis each test item receives, and the length of the test. Test item formats must be selected, performance standards set, and test items must be produced and entered into the test item bank.

### 1. Table of Specifications

The table of specifications is a blueprint, or plan, that clearly defines the scope and content of the test. With a clear understanding of what the test should cover, the test developer is less likely to let the scope of the examination be determined by the ease of item writing. Just as it is important that learning objectives be developed before instruction is planned, so too, it is necessary that the table of specifications be developed before the test is constructed. Both plant and training management should be involved in the establishment/approval of the test table of specifications.

There are many forms the table of specifications can take, and each form can have several variations. Two of these forms are presented.

#### a. Developing the table of specifications

If behaviorally stated learning objectives already exist, complete with action statements, conditions, and standards, the major portion of test planning has already been accomplished. What remains is to determine which objectives will be covered in the test, how many items will be included, and what the relative importance of the test items will be.

Figure 1 shows a table of specifications developed from a list of learning objectives. The objective statements indicate the type and level of performance expected of the trainee. But the instructor must select which ones will be tested on a given exam and establish the relative emphasis each instructional objective receives.

| Objectives for Training | Testing Emphasis (item weight) | Objectives to be Included in Test |
|---|---|---|
| I. Area A | | |
| 1. | 5% | yes |
| 2. | 10% | yes |
| 3. | 0% | no |
| 4. | 5% | yes |
| II. Area B | | |
| 1. | 10% | yes |
| 2. | 0% | no |
| 3. | 2% | yes |
| 4. | 0% | no |
| III. Area C | | |
| 1. | 5% | yes |
| 2. | 10% | yes |
| 3. | 2% | yes |
| . | | |
| . | | |
| . | | |

**Figure 1: TABLE OF SPECIFICATIONS FROM OBJECTIVES: This is an example of a table of specifications for a test created from a list of course objectives.**

As Figure 1 shows, Objective III.2 is given twice as much weight on the examination as III.1 and five times as much weight as III.3. These different weights should be based on the objectives' relative importance to success in the job environment and should reflect the relative time spent on the objectives in the training program. There are no hard-and-fast rules for determining the weights to be assigned to the various cells of the table of specifications. The test developer should obtain input from other trainers, subject matter experts, and line management and supplement this with his own prior experience. The trainees will expect the testing emphasis to be correlated with the emphasis stressed during training, and indeed this should be the case. Learning objectives can be assigned greater weights by using multiple questions or by assigning higher point values. Increasing the number of questions is the preferred way to increase emphasis.

Figure 1 further shows that objectives I.3, II.2, and II.4 do not appear in this examination. This is because they have already been covered in previous tests, will be covered in later tests, or can be tested in conjunction with other objectives. All learning objectives should be tested at some point during the training.

When detailed, behaviorally stated objectives are not available, or when it is impractical or impossible to obtain direct assessment of trainee performance objectives, a two-way table of specifications is normally developed. In this type of table the content area to be tested is identified, the number (or percentage) of items required is stated, and the types and levels of learning to be assessed are identified. The INPO GET-RP examination is a case in point. A generic test was developed for use across the industry without using the specific learning

objectives of individual training programs. Figure 2 is a two-way table of specifications for that examination program.

| TOPICS | LEARNING LEVELS | | | Total |
|---|---|---|---|---|
| | Knowledge A | Comprehension B | Application C | |
| 1. Fundamentals | 1 | 1 | 0 | 2 |
| 2. Biological Effects | 0 | 2 | 0 | 2 |
| 3. Administrative | 2 | 2 | 1 | 5 |
| 4. Exposure Control | 2 | 5 | 4 | 11 |
| 5. Contamination Control | 2 | 5 | 3 | 10 |
| 6. Monitoring | 1 | 3 | 1 | 5 |
| 7. Access Control | 3 | 3 | 1 | 7 |
| 8. Unusual Incidents/ Emergencies | 1 | 2 | 3 | 6 |
| 9. Protective Clothing and Respiratory Equipment | 0 | 1 | 1 | 2 |
| Total Number of Questions | 12 | 24 | 14 | 50 |

**Figure 2: TWO-WAY TABLE OF SPECIFICATIONS: Table of Specifications for INPO GET-RP Examination**

As Figure 2 indicates, the exam contains nine topic areas, three learning levels, and 50 questions. The numbers in each cell represent the number of questions used for the topic and corresponding level of learning. These numbers, which vary from 0 to 5, also represent the relative weight of each topic and level. For example, there are no knowledge-level biological effects questions (zero weight), while there are four application-level exposure control questions (8 percent of the exam).

How firm should the assignment of weights to each cell be? The weights should be a "best judgment" decision. They are usually established by the test developer with input from other instructional staff, plant and training supervisors, and job incumbents. Once established, they should remain relatively stable. The weights should not be changed to suit the test items developed. If the table of specifications calls for a large number of application-type questions, they should be developed. The reason many tests emphasize recall or factual

material is not because the instructor stressed only facts but because it is more difficult to write test items that measure the advanced mental processes of analysis, synthesis, and evaluation.

b. **Using the table of specifications**
When used to build the test, the table of specifications ensures the content validity of the examination. It is an effective method for keeping both instruction and testing on target. If distributed to the trainees, it can eliminate much confusion and misunderstanding. This is not "teaching the test" but rather "teaching for the test," a proper and efficient form of instruction.

2. **Selection of Test Item Format**

a. **Format types**
Before test items can be developed appropriate test item formats must be selected. There are four basic formats: oral, structured response, unstructured response, and performance.

*Oral:* Demands for accountability are not easily satisfied through the use of oral examinations. Examiner questions and trainee responses are generally not well recorded, and detailed, comprehensive answer keys are difficult to develop and use. However, oral questioning, an informal evaluation technique, is an excellent method for checking trainee understanding in the classroom, laboratory setting, or during on-the-job training. Oral questioning requires trainee participation, clarifies instruction, and motivates trainee learning. It is one of the fastest and most accurate methods to obtain feedback in the classroom setting. Oral questioning provides a two-way exchange between the trainee and the instructor, enabling confusing points to surface and to be clarified.

When oral exams, as opposed to oral questioning, are used, the test questions should be developed prior to administration, the acceptable answers must be recorded in advance, and the student responses must be graded and documented. Basic procedures to be followed for oral examinations should not differ significantly from those applicable to essay exams. There are two advantages to the oral exam. Oral exams do not require the examinee to record his answers in written form. This saves time on the part of the student; however, the burden is shifted to the instructor who has to both document the trainee response and record the evaluation. Oral examinations allow the instructor flexibility to add questions to the exam to test problem areas in more detail. It is important that these additional questions and their graded responses be documented. It is up to the training program personnel to decide when the advantages of oral exams outweigh the advantages of unstructured (essay type) exams. The obvious tradeoffs are flexibility versus ease of documentation.

*Structured Response:* Structured response test items are designed to limit the trainee's response to a small range of choices. Fill-in-the-blank is an example of a structured response item where the correct answer is a single term, phrase, or its equivalent. The vast majority of structured response items are of the variety commonly called objective test items. Objective test items normally take the form of alternate choice, multiple choice, or matching. Structured response test items are the most often used and misused. A frequent mistake is to use structured response test items when other methods, such as essay or

performance, are more suitable. The strengths of structured response items are that they allow reliable scoring, can be quickly answered, allow more material to be tested, and can be more easily analyzed and refined.

*Unstructured Response:* Essay and short answer items are the two principal types of unstructured response items. Unstructured response items are thought by many to be more effective than structured response items when testing for certain higher level skills such as analysis, problem solving, and critical thinking. This is not true. This thinking has come about primarily because it is easier to develop an unstructured response item to test higher skill levels. A look at any standardized scholastic aptitude tests will demonstrate that structured response items can also test higher level skills. The real value of unstructured response items lies in another area. Essay exams allow individuals to organize and express thoughts and concepts in their own words. Essay exams afford the trainee a greater opportunity to demonstrate individuality; a person can interpret the test item, respond accordingly, and have a unique view considered.

*Performance Tests:* While any test can be said to be a performance test (i.e., it measures some kind of performance), the term is generally restricted to those situations in which the trainee applies learning to doing actual or simulated job tasks. Asking a trainee to describe proper welding techniques is not a performance test; asking the trainee to create a proper weld is. When lesson objectives are concerned with direct job performance skills, the performance test is the preferred examination method. Time, expense, or equipment unavailability often preclude this test mode. Many utilities have found job performance measures to be an effective method of testing trainees' performance of job tasks. Like unstructured items, performance tests are subject to variation in scoring.

b. **Considerations in format selection**

There is no single best test item format for all situations. A format appropriate in one environment may be less appropriate in another. Each format has its advantages and disadvantages. Consider the following factors:

*Action Verb:* The most important factor in determining which test item format to use is the intent of the test item. What is it to measure? If the purpose is **list**, a short answer question is appropriate. If **identify** is an acceptable alternative, a multiple choice or matching item can be used. If the learning objective verb is **state**, oral questioning would be an appropriate technique. If a person must **compute** a value, an unstructured response word problem or a multiple choice item with suitable distracters could fit the intent of the learning objective.

Whenever possible, a "match" should be made between the learning objective and test item format. In cases in which this is not practical, the test item format should correspond as much as possible. It may also be that the objective is inappropriate for the training setting and should be changed. It is likely that the development of test items will cause some revision to the learning objectives.

*Facilities Available:* If time permits, the actual job environment may be used to perform the examination. In most training departments, training is divided into classroom, laboratory, simulator, and on-the-job instruction, with each training environment using the most appropriate test format.

*Number of Trainees Taking the Exam:* A key advantage of the structured response item is its quick scoring. If an exam is used for a large number of people, this may be the best choice.

*Time:* Essay examinations have a slight advantage when preparation time is limited. However, they generally need more time for scoring. Time is also a factor in the administration of examinations. It can easily take several hours to set up and administer a performance examination. Essay exams may require an hour to administer four questions, while four multiple choice questions can typically be completed in a few minutes.

3. **Use of References During the Exam**

Training environments use many teaching aids to promote trainee learning. Many of these aids (e.g., textbooks, manuals, and student guides) assist the trainee to master required subject matter. The trainee is expected to comprehend the material and commit important facts to memory. Once learned, the trainee should no longer need to refer to these materials. However, there are many references (e.g., tables, charts, schematics, and procedures) that trainees do not need to commit to memory but must be able to interpret and use on the job.

To test trainee use of these references, test items that demand that the references be used to solve a problem or reach a conclusion must be developed. This requires that the reference, or a sufficient subset of the reference, be provided during testing. This use of reference materials during testing has been referred to by some individuals as **open book** testing.

The test item developer must determine which references are necessary and how they will be used after reviewing the test objectives and the test table of specifications. While testing that includes the use of references is essentially no different than other written examinations, there are a few points to consider when using this method.

(a) References should be considered tools that the trainee uses to solve a problem. Do not directly test knowledge of the references (this should be done in a closed book section); rather, test for proper use of the references.

(b) Test the trainee's ability to locate, use, and apply the information found in the references.

(c) Keep the references and job aids made available during testing consistent with the conditions stated in the learning objectives.

(d) Familiarity with routine applications of a reference can lower the learning level of a test item to simple recall. Write test items that contain unique or varied circumstances that the trainee has not previously encountered. This makes the test item a true indicator of the trainee's ability to apply knowledge through the use of the references versus merely remembering an application from an earlier training session.

(e) Do not make a referenced test item easier or more difficult. Just because a test item has references provided, it is not appropriate to make the item more or less difficult than a comparable closed book item. While very difficult items may be useful in differentiating among the most able trainees, they are not appropriate for job qualification.

(f) Keep the referenced test items content valid. It is important that the test items address the use of references in a context similar to that found in the job environment. Keep test item requirements as close to real-life situations as possible. While providing a chart from a handbook is good, giving the trainee the handbook and requiring that the chart be found can be even better.

(g) Administer all closed book test items separately and before an open book test section. This ensures that the trainee does not find "giveaway" answers in the references.

(h) Allow sufficient time for the trainees to complete the test items. The more familiar trainees are with the references, the faster they can complete the items. However, be cautious that the test does not become a time test. Unless time is a crucial factor in the task, it should not be made a part of the test.

(i) Make the references and training aids that will be used during testing known to the trainees prior to testing. This is important for two reasons. First, if they are expected to bring the aids they must be notified, but more importantly, they need to know what will be expected during testing, i.e., they must use the references rather than memorize the references. It is a waste of effort to memorize facts that do not need to be memorized.

4. **Test Item Development**
Test items are generally developed early in the training development process. The TSD process recommends that test items be developed in the design stage following the preparation of learning objectives. The Principals of Training System Development Addendum I: Test Item Development provides guidance in the development of test items.

5. **Test Item Bank**
Training departments often keep test item files, or exam banks. Such files usually consist of previously used tests, answer keys, test items, and historical data on prior administrations. The experienced instructor understands the benefit of having this ready reference when called upon to develop an examination. Not only does it save a lot of time, but the resulting test items are significantly improved due to the modifications that normally accompany test administrations. Instructor training programs usually encourage instructors to maintain such a data bank and to collect test item analysis information on each use of a test item. This information may be collected and filed using a paper system or a computer system. The widespread use of computers has added significantly to the capabilities and flexibility of such test item storage systems. For example, multiple versions of an exam may be produced to increase test security during administration. Since most utility training organizations provide training by program area using several instructors, it is important that the test item data bank concept be applied at the program level. In this way, the size, scope, and uniformity of the testing process will be improved. Effective computer software can be used to increase the efficiency of test development, while providing an effective tool for test item evaluation and improvement. Some points to consider in establishing a test item bank are as follows:

(a) Establish the scope of the bank. Determine whether it will contain only objective-type test items or whether it will also include essay items, ones involving pictorials, graphs, etc.

(b) Establish effective security controls. This can be done by limiting access via restricted passwords. If the bank is on a separate computer system not generally accessible to plant personnel or trainees, the chance of unauthorized access is significantly reduced.

(c) Establish a program for test and test item analysis. Develop standard report formats that display the results of the test, test items, and each test item alternative. Include an option for compiling data on all administrations of each test item and use this information to help guide test revision and control test item quality.

(d) Consider using machine-scored answer sheets. This saves time, allows automatic input into the data base, and can be integrated with item analysis routines.

(e) Establish clear guidelines and procedures. Determine the required and authorized uses of the test item bank, then ensure that the test item bank is properly used and maintained.

(f) Consider installing a table of specifications program. A computer program that allows the instructor to create a table of specifications for each exam is a valuable tool. Working from a table of specifications is an effective way of controlling the content validity of a test.

(g) Develop a test item numbering system. It is likely that you will want to tie each test item to the following identifiers:

- program
- lesson plan
- objective
- test item type
- test item level
- point value

(h) Consider sharing information with other utilities. Sharing test item bank information can improve industrywide testing, save time, and maximize resources.

## C. TEST CONSTRUCTION

Following test planning and test item development, the test is constructed. Test construction requires that the test developer establish the test layout, assemble the test items, and write test directions.

### 1. Test Layout and Assembly

The test should be assembled in a format that is logical and easily understood by the trainee. It should follow conventional rules for ordering test items.

Written examinations should include typed or clearly written test items and should be reproduced so that each trainee has an examination. Writing the questions on the board or stating the questions orally is inviting misunderstanding. An oral examination is not meant to be a written test given orally but is a unique situation requiring two-way communication.

The test should be clearly labeled. The course, test title, associated unit of study, administration date, and test form should be stated on the exam. If the examination is to have trainee responses written on the examination, it is a good idea to put this identifying information on a cover page where the trainee's name, employee number, or other required information are entered. The following arrangement of test items is preferred:

(a) first, group all items using a common body of information (e.g., diagram, table, or scenario) even if test item formats must be mixed,

(b) then, group all items of the same format,

(c) then, group all items dealing with the same objective, and

(d) finally, group items from least to most difficult

Some examinations consist of only one format, but most instructor-developed exams contain a variety of item types. While using only one format (typically, multiple choice) has the advantage of simplicity and clarity in giving only one set of directions, it is more difficult and time consuming for the test developer to force all questions into one format. There is nothing wrong with a variety of formats. However, to keep the exam responses ordered from simple to complex, the following order of test items is suggested.

(a) True/false or alternative response items

(b) Multiple choice items

(c) Matching items

(d) Short answer items

(e) Essay questions

When a diagram, drawing, or other block of information is used with a test item or items, it should be placed above the item stem or test item if possible. Also care should be taken when splitting material between pages. Avoid splitting a test item, but if one is split, present all of the item alternatives on the same page. Keep matching items together on the same page.

**2. Test Directions**

Each examination should have clear written directions. These directions should tell the test taker what to do, how to do it, and how to record the responses. General directions should be given for the total test, with specific directions given for each section, subpart, and item type. While the test administrator should orally present the directions, the written directions should be clear enough to enable the trainees to complete the test without any further instructions.

Trainees should be told the value of test items and how they will be scored. The trainee should know whether partial credit will be given, what degree of precision is required, whether units must be identified (such as psi, ohms, rems), and, for calculations, if work must be shown. Time limits should be stated.

When developing the instructions, keep them succinct. Make important points stand out by using a different size or *style of type* or by underlining. Have others review the directions to check for inconsistencies or potential misunderstandings. Consider including sample items with the directions when introducing difficult or unusual item types. Clear directions will help maintain the reliability and validity of the test. Figure 3 provides sample test directions.

DIRECTIONS

This is a test of your radiological protection knowledge. If you score 80% or better on this exam, you will be exempted from generic radiological protection training.

Make no marks on the exam booklet. Place all answers on the answer sheet. Use scratch paper or the back of your answer sheet for any calculations.

Decide which is the best answer from among the alternatives, then mark the appropriate space on the answer sheet. Your answer sheet will be scored mechanically, so it is very important that you mark your answers correctly.

1. Mark only one space for each question on the answer sheet.
2. Use only a number 2 lead pencil on the answer sheet.
3. Make sure your mark fills the space, but does not go outside the space.
4. If you change your mind, erase your first mark completely and make another mark.
5. Keep your answer sheet clean; stray marks may be counted as errors.
6. Since all unmarked questions will be counted as wrong, answer all questions even if you are uncertain which answer is correct.

If you have any questions, ask the exam proctor now.

You have 45 minutes to complete this exam. If you finish early, check your work. Be sure that you have answered all the questions.

You may begin.

**Figure 3: EXAMPLE DIRECTIONS**

## D. TEST ADMINISTRATION

Improper administration of an examination can have an adverse effect on the usefulness of the test results. Many psychological, intelligence, and personality tests require that detailed procedures be followed and that the test administrator be trained and certified to administer the exam. A standardized achievement test also requires that specific administration procedures be followed.

Locally developed examinations also require controlled administration. The test administrator must ensure that a suitable environment is established, that test directions are given, and that there is proper supervision.

1. **Establish Environment**
   An effective testing environment requires that attention be paid to the physical qualities of the test setting and the emotional climate in which the trainee must perform. Noise, poor lighting, lack of ventilation, excessive heat or cold, and frequent interruptions will lower trainee test performance. The test administrator should maximize, to the extent possible, the conditions for testing. This may be as simple as scheduling testing in the morning if the classroom becomes too hot in the afternoon.

   While most instructors and test administrators are aware of the physical testing environment, many do not give sufficient consideration to the emotional environment that they establish. Inappropriate attitudes range from "don't worry about this exam, it doesn't really measure your trouble-shooting skill" to the overbearing, "nothing in your academic past will challenge you like I will, not many people pass my course!" The testing environment should be conducive to effective testing, just as the classroom environment should be conducive to effective learning. A good emotional climate is important in building motivation, reducing anxiety, and improving communications. Consider the following points.

   - Make the purpose of the test clear. All trainees know that a final exam is an evaluation. They may not know that many tests can be diagnostic or learning experiences.
   - Emphasize the need for accurate test results. Trainees need to know that the instructor expects the best possible trainee performance on all tests and that training and career decisions cannot be effectively made without the conscious effort of the trainee.
   - Minimize test-taking anxiety. Put the test in proper perspective. The test score is only one of several indicators that the instructor collects over the course of time. If the trainee has prepared adequately, test results will reflect it.

2. **Provide Directions**
   Effective written directions are sufficient to guide most trainees. However, no matter how clear or precise, some trainees will still misread or misunderstand the directions. Every test administration should begin by orally reviewing the test directions and clarifying any misconceptions. If the test item types are new to the trainees, a sample item can be discussed before beginning the exam. Particular attention should be given to trainees with special needs. Once the test has begun, it is a good idea to move around the classroom, looking over the trainees' work, to ensure that everyone is following the directions.

3. **Monitor Exam**
   It is extremely important that test results provide an accurate indication of a trainee's performance. Exam scores can have significant meaning since they can affect job placement, promotion, job security, and salary. No one is served if a trainee performs well or poorly for the wrong reason. Both the abilities of the trainee and the effectiveness of the course and instructor are misrepresented. Effective exam proctoring can put a misguided trainee back on track. It can also prevent cheating. Unfortunately, some instructors believe that just being in the room during test administration is sufficient. Some instructors do not even believe this is essential.

Training procedures should provide definitive guidance for exam monitoring. The instructor should realize that even the best trainees can occasionally misinterpret the directions, and, without adequate supervision, pressures may tempt some trainees to cheat.

Several techniques can deter trainees from cheating. The single best way to is to observe trainees carefully during testing. Some training department procedures require that each trainee sign an affidavit that the work is solely the individual's. This has some deterrent value. However, it should not be allowed to replace other useful methods. These include spacing trainees during testing, using multiple exam forms, and revising the exam for each testing session. A clear policy on academic honesty should be established and enforced.

## E. SCORING THE EXAMINATION

Exam scoring will vary, depending on the purpose of the exam. This section addresses the various methods of scoring and provides some suggestions on using them.

### 1. Self-scoring

Self-scoring is often used for tests where the results will not be collected by the instructor. These tests are primarily self-instructional; they inform trainees of their current ability. Self-scoring is also useful for personality, interest, or career-planning inventories. Answers can be provided at the end of the examination, or a variety of techniques can be used to disclose the correct responses. A variation on self-scoring is to have trainees exchange papers and score them in class. This saves the instructor time and can provide immediate feedback for both the instructor and trainee.

### 2. Hand Scoring

Hand scoring is the most common scoring technique. Usually a scoring key is created on a strip of paper that may be placed next to the test form, or a blank test form is completed with the correct answers. For multiple choice test items, separate answer sheets can be used. An answer key can then be created by punching out the correct answers. The resulting overlay allows rapid scoring. The overlay should be made of a transparent material, such as an overhead transparency, so that the instructor can easily detect omitted or multiple responses. The scoring of essay items is discussed under unstructured test items.

### 3. Machine Scoring

When a large number of structured response exams are to be scored, machine scoring may be indicated. In addition to saving time, the ability to enter the results directly into a computer system provides many other benefits. Trainee records can be updated, test analysis data can be automatically computed to aid in test refinement and program evaluation, and reports and records can be produced easily once the initial programming is complete.

### 4. Scoring Unstructured Test Items

Many examinations are of the unstructured response type. These examinations cannot be machine scored but must be reviewed individually by the instructor. Because of this, scoring unstructured response questions poses some unique challenges. It takes diligence on the part of the instructor to prevent these test items from becoming "subjective" test items.

To minimize the subjectivity in scoring any unstructured response item, several guidelines should be followed.

(a) Compare the answer key to several actual exam responses. Some trainees may take a different approach than the answer key anticipated, and a refinement to the standard may be necessary. If the standard is changed, all papers should then be rescored using the revised standard.

(b) Periodically review the answer key. It is easy for an instructor's standard to drift as several exams are scored. To guard against this the instructor should periodically review the answer key. Also, by occasionally reviewing those items scored earlier, the instructor can confirm that the standards are being applied consistently. Even with these measures, some inconsistency is inevitable. One problem is that of an item response following several good or several poor responses. The tendency is to score the item low if it follows several high scores, or score the item high if it follows several low ones. Shuffling the exams between review of exam items, while not eliminating the problem, allows these effects to be offset by random sequencing.

(c) Score each item separately. Each test item should be scored for all exams before the next item is scored. Scoring one item at a time allows the instructor to concentrate on just one standard. This increases consistency when assigning points or categorizing items.

(d) Avoid interruptions while scoring a set of responses. The bias an instructor has toward an essay item may change from one time to another. If a bias exists it should be consistently applied to the responses of all trainees. For example, an instructor may be irritated one afternoon and alert the next morning. By scoring all response sets at one sitting, if a bias exists, its effects on trainee scores will be consistent.

(e) Provide comments and make corrections. A trainee who does not receive full credit for an answer will want to know why. Appropriate comments can explain the score received. For trainees to learn from their mistakes, they must be told what errors were made and how to correct them. Another value in providing comments is the ability to tally the various comments and analyze the results for test item improvement.

# CHAPTER 4

## A. MEASUREMENT CONCEPTS

To develop effective examinations and analyze test results, the instructor should have a knowledge of basic measurement concepts. While most instructors and test developers will not be required to perform complicated statistical analyses, an understanding of some basic concepts is beneficial in interpreting and refining the testing process.

### 1. Reliability

Reliability is the ability to give consistent results. Reliability is functionally defined as the consistency between two separate measurements of the same thing. If a test gave perfectly consistent results, it would be perfectly reliable and would have a reliability coefficient of 1.00. Conversely, a test with no reliability would have a reliability coefficient of 0.00. No testing situation is perfectly reliable.

Ebel and Frisbee[4] provide the following definition:

> *The reliability coefficient for a set of scores from a group of examinees is the coefficient of correlation between that set of scores and another set of scores on an equivalent test obtained independently from the members of the same group.*

This definition makes several points. First, it states that reliability is functionally defined in terms of a correlation coefficient. Several coefficients can be used, depending upon the nature of the data; the most common is the Pearson Product-Moment Correlation Coefficient.[5] Second, it states that a group of examinees must be tested to determine reliability. Third, the group must have two sets of scores taken independently. A correlation can only be computed using group data. The relationship between the two measures will determine the value of the reliability coefficient.

A test would be perfectly reliable if there were no error in measurement. In reality though, every test administration includes some error. Therefore, each test score is considered to be the sum of a true score component and an error score component. As the size of the error component decreases in proportion to the size of the true score, the reliability increases. If all sources of error could be eliminated, reliability would reach the theoretical maximum of 1.00.

Sources of error are practically limitless. They range from poor test items (ambiguous items, multiple answers, and misstated items), to adverse testing conditions (noise, heat, poor lighting, interruptions, limited time, and complicated answer sheets), to examinee readiness (lack of sleep, mental attitude, alertness, and physical condition), and scoring errors (improperly scored

---

[4]Ebel, R. L., and Frisbie, D. A., Essentials of Educational Measurement, 4th ed. (Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1986).

[5]For a detailed explanation of the correlation coefficient see Glass, G. V., and Stanley, J. C., Statistical Methods in Education and Psychology (Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1970).

items, incorrect answer keys, and biased essay item scoring). Anything that contributes to measurement error reduces test reliability. There are three common methods of statistically estimating reliability: measures of stability, measures of equivalence, and measures of internal consistency.

**a. Measures of stability**

Measures of stability, sometimes referred to as test-retest reliability, measure examinee error. This reliability is estimated by administrating the same exam to the same group of people at two points in time (e.g., morning and afternoon or on successive days). Any learning that takes place between successive administrations confounds the issue and lowers the reliability estimate. Collecting this type of reliability information is not recommended as a routine procedure in a training program.

**b. Measures of equivalency**

Another way to establish reliability is through the administration of equivalent forms of a test. The correlation between the scores obtained on the two forms represents the reliability coefficient of the test. Equivalent-form reliability requires truly parallel alternate forms. The two tests should contain the same number of items; the items should be in the same form, cover the same content, and be at the same level of difficulty.

**c. Measures of internal consistency**

A more usable measure than test-retest or equivalent-forms reliability is found in several measures of internal consistency. These methods allow a reliability estimate to be made from a single test administration. One method, split-halves reliability, divides the test into two equivalent halves and uses these subtests to calculate reliability. Several other internal analysis methods have been developed. Kuder and Richardson developed the formulas referred to as KR-20 and KR-21. They are applicable to tests scored dichotomously, that is, each item is either right or wrong.[6] For essay tests, or tests scored other than right or wrong, Cronbach's coefficient alpha is an applicable formula.[7] A basic measurement text can provide the necessary formulas and information to calculate these and other measures of reliability.

There are many factors that influence test reliability. The instructor should be aware of how these factors affect reliability, even though they cannot all be controlled. The reliability coefficient will be increased by each of the following:

---

[6]For a more complete discussion of reliability estimates and their formulas see Mehrens, W. A., and Lehmann, I. J., Measurement and Evaluation in Education and Psychology, 3rd ed. (New York, N. Y.: Holt, Rinehart and Winston, 1984) pages 275-277.

[7]Cronbach, L. K., "Coefficient Alpha and the Internal Structure of Tests," Psychometrika, Vol. 16, pages 297-334, 1951.

(1) Increasing the length of the test, the simplest way to increase reliability (see Figure 4)[8]

(2) Increasing the homogeneity of the test items, i.e., make all the test items similar in what they test

(3) Using test items that are more discriminate, i.e., items that separate the unknowledgeable examinee from the knowledgeable examinee--This is the most effective way to increase reliability.

(4) Using test items that are of medium difficulty rather than extremely easy or hard ones

(5) Administering the exam to a group of examinees with a wide range of ability regarding the subject tested

| Number of Test Items | Reliability |
|---|---|
| 5 | 0.20 |
| 10 | 0.33 |
| 20 | 0.50 |
| 40 | 0.67 |
| 80 | 0.80 |
| 160 | 0.89 |
| 320 | 0.94 |
| 640 | 0.97 |
| ∞ | 1.00 |

**Figure 4: TEST LENGTH VERSUS RELIABILITY: Effects of successive doubling of the length of an original five-item test, the reliability of which is assumed to be 0.20.**

**d. Reliability of criterion-referenced tests**

The procedures that have been discussed were developed to determine reliability coefficients for norm-referenced tests. However, they are applicable to criterion-referenced tests when there is variability in the test scores. Since the variability of criterion-referenced tests is likely to be less than that of norm-referenced tests, the reliability coefficients derived will be lower; that is, they will be conservative estimates of test reliability. Some proponents of criterion-referenced measurement profess to be only interested in the accuracy of the

---

[8] Ebel, R. L., and Frisbie, D. A., Essentials of Educational Measurement, 4th ed. (Englewood Cliffs, N. J.: Prentice Hall, 1986), page 83.

decision (pass or fail). For those who insist on only using a pass-fail interpretation of criterion-referenced tests, there are other formulas that can be used.[9]

2. **Validity**

Validity is the most important characteristic of any test. To be valid, a test must measure what it is intended to measure. A test can be reliable but not valid. However, it is impossible for a test to be valid and not reliable. A paper and pencil test can be reliable in measuring knowledge of certain welding fundamentals but not valid for measuring welding skill. Similarly, a performance examination that measures welding skill would not be valid if the instructor did not adhere to clearly defined standards when evaluating trainee performance. If the exam standards are not consistently followed, the exam is not reliable, and if it is not reliable, it cannot be valid. As these examples indicate, validity is not solely a characteristic of the test, but rather it is a characteristic of the test and the intended use of that test. A test is only valid or invalid for some given purpose. Test data is used to make decisions. If the test data enables good decisions to be made, the test is valid for that purpose. Therefore, tests are predictors of future events; the accuracy of the prediction defines the validity of the test.

A discussion of validity comes down to the question of how validity can be assessed. Just as we cannot know the future, we can only collect evidence that certain events will occur. Evidence supporting test validity is generally grouped into one of three categories: content evidence, criterion-related evidence, and construct evidence.

a. **Content validity evidence**

To establish content validity, it must be shown that the test items sample the domain being predicted. This requires that the content domain be clearly defined. The content domain is defined by training program goals and course objectives. It is also defined by task performance, that is, those skills and knowledges that a trainee must demonstrate after instruction.

Content validity evidence is established by comparing the responses to test items with the expected responses. The establishment of content validity means that the objectives of training must be clearly defined and the test results must accurately demonstrate those objectives. If a job skill requires quadratic equations to be solved, content validity evidence would require test items that result in solved quadratic equations. The comparison of learning objectives to test results must be done by persons knowledgeable in the content area. There is no commonly accepted mathematical formula for measuring content validity.

A detailed table of specifications and/or a complete listing of learning objectives are the first steps in the establishment of content validity. The next step is to review the content to ensure an appropriate content-to-objectives match. The last step is to ensure that the test is appropriate for the trainee population to be tested.

---

[9] For a review of criterion-referenced indices see Kane, M. T., and Brennan, R. L., "Agreement Coefficients as Indices of Dependability for Domain-referenced Tests," Applied Psychological Measurement, Vol. 4, pages 105-26, 1980.

b. **Criterion-related validity evidence**
To establish a test's criterion-related validity, the test's results are compared to another measurement, the criterion, which has previously been accepted as valid. If an essay examination has been used and accepted as valid for a course final examination, but takes three hours to administer and 20 hours to score, a new 50-item multiple choice exam, which can be administered in two hours and graded in 20 minutes, could be a great improvement if it can be proved to be valid. The easiest way to establish the new test's validity is to compare its results to the previous exam and show that the results of the two tests are the same. This is done by establishing the correlation between the two exams, using one of the common correlation coefficients.

The type of criterion-related validity just discussed may also be referred to as **concurrent** validity because the criterion measure, the essay exam, is measured at or about the same time as the newly developed multiple choice exam. If the criterion measure is a performance that occurs in the future, then **predictive** validity is being established. Most training program examinations are given with the hope that they will predict job performance. Predictive validity can be established by comparing the examination results to subsequent job performance indicators like task performance, supervisor evaluations, job tenure, or job advancement. However, if an appropriate criteria is not available, the predictive validity correlation coefficient may be very low. The criterion measure should be carefully selected and thoroughly described. The group for which the validity is being established must also be clearly defined. An examination that is valid for apprentice mechanics may be less valid for journeyman mechanics. It is important to remember that validity measures are specific to the unique characteristics of the group being tested.

c. **Construct validity evidence**
A test is said to have construct validity when the test scores vary as the theory underlying the construct would predict. While, in theory, every test can be said to be based on some construct, construct validity is of primary importance in the areas of research and psychological evaluation. For most ability and performance testing the constructs are obvious and seldom need be addressed. For example the construct underlying a welding performance test is that a sequence of sample welds performed during an examination is an accurate predictor of the quality of welding a worker will perform on the job. Barring outside factors such as worker health, management, and working conditions, most people would accept the construct validity of the welding test.

3. **Standard Error of Measurement**
The standard error of measurement provides an estimate of the accuracy of a **single** test score. Once the test's reliability and standard deviation have been determined, the standard error of measurement can be estimated. It allows the test user to evaluate how much an individual's test score might vary from a true score. Using the standard error of measurement, a probability estimate can be established based on the observed score. An illustration will show how this is done. If a 100-point exam has been administered to 50 trainees, we can score the exams and calculate the test mean (arithmetic average) and standard deviation.

If we have a reliability coefficient (r), we can then estimate the standard error of measurement using the formula: $SEM = SD\sqrt{1-r}$

Where: $SEM$ = standard error of measurement
$SD$ = standard deviation of the test
$r$ = correlation coefficient of the test

If SD=5 and r=0.84 (a fairly low standard deviation and a reasonably high correlation coefficient for a classroom examination), the resulting standard error of measurement is 2. The standard error of measurement reflects the variability in the score due to error and, assuming that measurement error is normally distributed, can be interpreted in terms of the normal curve frequencies. Approximately 68 percent of the cases fall within plus or minus one standard error, approximate 95 percent of the cases fall within plus or minus two standard errors, and approximately 99 percent of the cases fall within plus or minus three standard errors. Taking this into account, for a trainee who scored 81 on an exam, we can be 68 percent confident that the true score lies somewhere between 79 and 83. We are 95 percent confident that the true score is between 77 and 85, and we can be 99 percent confident that his true score is between 75 and 87.

If the standard deviation of the test equals 7 and the correlation coefficient of the test equals 0.49, the resulting standard error of measurement would equal 5, which would not be unusual for an instructor developed exam, and the confidence intervals would become 76 to 86 for a 68 percent confidence level, 71 to 91 for a 95 percent confidence level, and 66 to 96 for a 99 percent confidence level. Thus, if the exam had a passing, or cutoff score, of 80, a trainee with a measured score of 81 could have a true score that fell well below or above 80.

Considering the above information it is easy to see that an instructor must be cautious in interpreting test scores, especially if the test exhibits low reliability or a wide standard deviation. Only when the test scores are well above, or well below, the cutoff score can we be confident that the trainee has clearly passed or failed the exam.[10]

An instructor should continually strive to improve test reliability and thus reduce the standard error of measurement. However, even the best test may incorrectly pass or fail a trainee. The wise instructor does not allow a single test to determine whether a trainee passes or fails. Rather, a pass/fail decision is made only after obtaining several performance measures.

4. **Norm-referenced Versus Criterion-referenced Testing**
A test is a measurement tool for collecting information. To establish the meaning of the information collected, some type of reference system must be used. The two reference systems generally used are known as norm-referenced and criterion-referenced.

Trainers should know when to use each of these reference systems. When it is important to **differentiate among individuals**, norm-referenced test interpretations should be made. If, however, we want to know that **a person has achieved a specific set of objectives** then

---

[10]For a more complete discussion of SEM see: Anastasi, A., Psychological Testing, 3rd ed. (Toronto, Ontario: The Macmillan Company, 1968), pages 94-95.

criterion-referenced interpretations are appropriate. Often overlooked is the fact that norm-referenced and criterion-referenced tests may be very similar, if not identical, in content and format. Norm-referenced tests differ, however, from the criterion-referenced tests in their attempt to spread the range of test scores and hence discriminate between trainees. Both test types should be made content valid, and both should be based on clearly defined objectives.

In many training situations the test user is not interested in a norm-referenced test interpretation. In these situations an accept/reject or pass/fail decision is desired. An example of this situation would be a state driver's license examination in which the examinee either meets or does not meet the minimum criteria. However, those using criterion-referenced testing must be aware of its limitations.

(a) Criterion-referenced tests have no more inherent content validity than do norm-referenced tests.

(b) Criterion-referenced testing can measure essentials but may fail to encourage maximum development.

(c) Questions of test reliability, involving test length, and the problem of setting cutoff scores can only be answered by reasonably lengthy and complicated procedures that require specific decisions regarding the costs of making errors.

(d) Criterion-referenced tests, like all tests, must be interpreted cautiously. If trainees fail to master an objective, the fault may be with the trainee, the instruction, the test items, the standard, or the objective itself.

Both criterion-referenced and norm-referenced interpretations have their place in evaluation. To summarize, one should use criterion-referenced interpretation when

(a) The interest is mastery, and the trainee either has or has not mastered the objectives.

(b) There is sound evidence that a specific cutoff score can be validly established, and the knowledge demonstrated above or below the cutoff score is not important.

(c) Program evaluation requires determining which trainees meet which objectives.

(d) Diagnosis of an individual trainee's specific learning deficiencies is desired.

One should use norm-referenced interpretation when

(a) It is desired to measure performance beyond required competency.

(b) It is desired to rank order all trainees in a program or on an exam.

(c) A selection decision is being made, and the best or poorest performing trainees must be selected (generally a fixed-quota situation).

(d) Comparison with other groups, locally, regionally, or nationally is desired.

Both criterion-referenced and norm-referenced measurement are useful for the interpretation of test results. Each can answer specific questions that are of interest to the test user.

## B. TEST ITEM EVALUATION

Test item evaluation can identify problems with a test item and suggest possible improvements. It can identify information that should be reviewed with trainees following test administration and can also be used to improve instructor test development skills.

### 1. Item Analysis

Item analysis is the process of reviewing each trainee response to assess the difficulty and discriminating ability of each test item and the effectiveness of each alternative. Each of these aspects will be discussed.

#### a. Item difficulty index

Test item difficulty is measured by dividing the number of trainees who answered the item correctly by the total number of trainees who were tested. This value, expressed in a percentage, can range from .00 if no one answers an item correctly, to 1.00 if everyone answers it correctly. Easy items have high indices near 1.00, and difficult items have low indices near 0.00.

#### b. Item discrimination index

Item discrimination refers to the ability of the test item to differentiate between high test scorers and low test scorers. Some assumptions are crucial to understanding this index. The first assumption is that the overall test is a valid measure of the knowledge area being tested; that is, knowledgeable trainees will achieve high overall scores, and uninformed trainees will achieve low scores. If we compare the scores on an individual item with the total test, we would expect that the better prepared trainees would answer the item correctly most of the time while the less prepared trainees would answer it correctly less often. This ability of an item to distinguish between the two types of trainees is referred to as its discrimination index.

The discrimination index is calculated by first selecting a high and low group, usually the top and bottom 27 percent of the tested group. For each test item, the item discrimination index is calculated by subtracting the number of trainees in the lower group who answered the item correctly from the number in the upper group who answered correctly, and dividing this result by the number of trainees in either group.

$$\text{Discrimination} = \frac{\text{Right (Upper)} - \text{Right (Lower)}}{\text{Number in Upper (or Lower)}}$$

The value is expressed as a decimal percentage and can range from -1.00 to 1.00. A positive value indicates a positive discrimination, and a negative value indicates a negative discrimination. Item difficulty affects discrimination. If an item is too easy or too hard, the discrimination index will be lowered. To maximize item discrimination values, the test item difficulty should be midway between the likely chance score and the maximum possible

score. What is a good value for discrimination? For instructor-developed classroom tests the following values are suggested.[11]

| Index of Discrimination | Item Goodness |
|---|---|
| 0.40 and up | Very good item |
| 0.30 to 0.39 | Reasonably good but subject to improvement |
| 0.20 to 0.29 | Marginal item, usually capable of improvement |
| Below 0.19 | Poor item, should be replaced or improved |

**Figure 5: DISCRIMINATION VALUES**

c. **Alternative Analysis**

Alternative analysis is the process of reviewing each test item alternative (correct answer and distracters) to identify how many trainees from the upper and lower groups chose each alternative. Alternative analysis can be made by calculating an index for each alternative, as the discrimination index was calculated for each item. Or an analysis can be done by visually reviewing the alternative results. The results can then be evaluated for potential change to the test item. Figure 6 provides an example to illustrate this process.

| | Alternatives | | | | | |
|---|---|---|---|---|---|---|
| | A | B | C | D* | OMIT | |
| Upper 27% | 0 | 3 | 1 | 23 | 0 | Diff=.70 |
| Lower 27% | 0 | 2 | 10 | 15 | 0 | Disc=.30 |
| Total | 0 | 5 | 11 | 38 | 0 | |
| * correct response | | | | | | |

**Figure 6: ALTERNATIVE ANALYSIS**

The above test item administered to 100 trainees has a difficulty index of .70, a reasonably good value. The discrimination value of .30 is also good.[12] However, through the alternative analysis process we see that distracter A is not being picked by any trainees. This suggests that it could

---

[11] Ebel, R. L., and Frisbie, D. A., Essentials of Educational Measurement, 3rd ed. (Englewood Cliffs, N. J.: Prentice-Hall, 1986).

[12] The ideal difficulty value for an item to provide maximum discrimination is halfway between the average chance score and the maximum possible score. 0.60 is the ideal value for a five item multiple choice question.

be replaced by a more plausible alternative. A better distracter for A would probably raise the discrimination value for this item. The change might or might not increase the difficulty of the item, depending on how the other choices are affected. The only way to see how the new item preforms is to administer the item to a new group of trainees and conduct alternative analysis on the new item.

A look at alternative B shows that one more high-scoring student chose this distracter than low-scoring students. While this is not the desired trend, the difference is too small to be significant. This small difference is attributed to chance.

3. **Use of Item Analysis Information**
Item analysis data is useful when reviewing and improving test items; however, it must be used with caution.

(a) Item discrimination and item validity are not synonymous. Validity addresses whether the item measures what it is intended to measure. The criterion is usually some external performance. In calculating discrimination, an internal criterion, total test performance is used. A valid item will usually discriminate, but a discriminating item is not necessarily valid.

(b) Item discrimination is not a perfect indicator of item quality. It has already been mentioned that a very easy or very difficult test item will not discriminate effectively. However, in many testing situations, primarily in criterion-referenced testing, many items will display a high difficulty value (above .80) indicating an easy item. Another factor that can lower item discrimination is the specificity of the test items. Recall that the basic assumption of item analysis is that the knowledgeable trainee will score high, and the less informed trainee will score low on all test items. For example, an exam covering metalurgical theory and welding practice administered to a group including engineers and mechanical maintenance personnel might show some undesired results due to the variety of test objectives and the variability in trainee skill and knowledge.

If the test were heavily weighted toward theory, the engineers would more likely be the "knowledgeable trainees," but since the mechanics would likely answer more practical test items correctly, these items would exhibit negative discrimination values. The solution to this dilemma is not to try and modify the negative descriminating test items, but rather to make the test more homogeneous. Place the theoretical and practical objectives into separate tests or place the trainees in separate groups.

(c) Item analysis data varies. Item analysis data is extremely variable for small samples. For large samples the data is still influenced by the nature of the group being tested and the teaching techniques that were employed. Difficulty and discrimination data should be considered after it is determined that an important instructional objective is being measured. Good item analysis data values are inconsequential if the appropriate objectives are not being measured. An item that is clear, technically correct, and discriminates positively should be used until it can be improved or replaced with a better item. An item that discriminates negatively, unless based on a small sample, should be reviewed. If an item review and alternative analysis do not uncover any noticeable problems, then the item can be used until a larger sample is reviewed. Random error may have caused the anomaly.

(d) Items should never be selected solely on item analysis data. Test items should be selected based on the relevance of the items to the learning objectives being tested. Item analysis should be used to choose the best items among those that fit the objectives and as a tool to improve the items. Items should never be selected solely on their statistical properties.

4. **Item Analysis for Criterion-referenced Tests**

There is disagreement among testing experts on whether the traditional item analysis procedures described above, which are designed for norm-referenced tests, can be applied to criterion-referenced tests. Some argue that the basic assumption of variability across the tested group is so narrow as to invalidate the process. In response to this, some researchers have proposed discrimination indices based on pre- and post-tests. These indices are effective but few training settings use pre- and post-tests, thereby making this process impractical. However, most criterion-referenced tests still exhibit sufficient variability to make the item analysis procedures useful. Difficulty indices can be expected to be much higher, in the range of .70 to 1.00, indicating easier items. This naturally lowers the discrimination values such that many good items will have near zero discrimination indices. However, negative discrimination indices are still significant, indicating that the items may be ineffective and need replacing. Also, alternative response analysis can highlight weaknesses in the item alternatives whether the test is criterion-referenced or norm-referenced.

# CHAPTER 5

## A. INTERPRETING TEST SCORES

As stated earlier, obtaining a valid test score is a major step in the evaluation process. This step is called measurement. The next important step is to interpret the meaning of the test score. What does a raw score of 35 mean? What does a raw score of 66 mean? For these scores to have meaning they must be referenced to a standard or criterion. Two referencing systems are commonly understood, norm-referenced and criterion-referenced. Not all measurement specialists agree on what these reference systems mean or how they should be used. Ebel and Frisbie have concluded that norm-referenced and criterion-referenced terminology has been used with such varied meanings that a redefinition of terms is necessary. They provide an informative discussion of the various types of test-score interpretation. Figure 7 presents their categorization.

| I. Content-referenced | II. Group-referenced | III. Criterion-referenced |
|---|---|---|
| A. Domain-referenced | A. Norm-referenced | A. Content-referenced base |
| B. Objectives-referenced | B. Treatment-referenced | B. Norm-referenced base |

**Figure 7: CATEGORIES OF TEST SCORE INTERPRETATION**

Content-referenced test interpretations are made when a trainee's performance level is compared to a defined set of knowledge and skills. The use of instructional objectives results in a special type of content-referenced interpretation termed **objectives-referenced**. In this referencing system test items are written to correspond with each instructional objective. Scores are interpreted in terms of achieving the objectives. For a group of objectives, the interpretation may be directed to achievement or non-achievement of each objective or to the proportion of objectives achieved. **Domain-referenced** interpretations are made when test items represent only a sample of the knowledge and skills of interest, that is, when not all objectives or all knowledge or skills are tested.

Group-referenced interpretations are made when a trainee's score is compared with other individuals in a specific group. Norm-referenced test interpretation relies upon a large number of scores that comprise the reference group. Treatment-referenced interpretation occurs mostly in research studies. The results of one treated group are compared to another group having no treatment or a different treatment.

Criterion-referenced interpretations are made when a trainee's score is compared with a cutoff score that represents a performance standard. Those at, or above, the cutoff score pass: those below fail. Trainees are not compared to each other, and no analysis is made to establish which objectives in the domain of interest were achieved. The establishment of a performance cutoff score can be either content-referenced or group-referenced.

Of the referencing systems discussed, each one has its proponents, and there is certainly merit in each system. Depending on the purpose of testing, however, one method may be more appropriate than another. Many situations can benefit from a blending of the methods. Ebel and Frisbie note that the terms and relationships relating to referencing systems have not been used uniformly

by measurement specialists. They have distinguished between content-referenced and criterion-referenced interpretation.

> *. . . in an attempt to highlight and to eradicate the popular misconception that the mere use of cutoff scores assures absolute score interpretation. The fact that cutoff scores may be established with either a group referenced or a content referenced basis seems to be overlooked too frequently. It is true that a cutoff score can be set for any test, but it is also true that we cannot satisfactorily interpret the scores from such a test unless the basis for establishing the cutoff score is known.*[13]

1. **Norm-referenced Interpretation**
   Norm-referenced interpretations require that group data be reviewed, described, and synthesized to serve as a point of reference. This is done by using descriptive statistics. Measures of central tendency and variability are typically calculated. Individual scores can then be compared to group data. To understand norm-referenced interpretations, the instructor must be able to interpret the necessary statistics. Many instructors are familiar with these basic concepts; for those who are not, an elementary statistics textbook can provide instruction or review. Concepts that should be understood include frequency distributions, histograms, frequency polygons, shapes of data distributions, and measures of central tendency, variability, and relationship.

   a. **Arranging data**
      A basic method of arranging data is to place raw test scores in ascending or descending order. Once ordered, the instructor can gain some understanding of the range and distribution of scores. An even better way of looking at a score distribution is to develop a frequency distribution, histogram, or frequency polygon. Figure 8 shows a frequency polygon for 30 mechanical maintenance exam scores.

      The polygon takes on a shape that defines the characteristics of that group of scores. This shape is referred to as the curve. The distribution shown in Figure 8 is typical of that exhibited in a performance-based exam using an 80 percent cutoff score. Each group tested will have its own unique curve. However, several curve characteristics occur with such frequency that they have been discussed and named.

---

[13] Ebel, R. L., and Frisbie, D. A., Essentials of Educational Measurement, 4th ed. (Englewood Cliffs, N. J.: Prentice Hall, 1986).

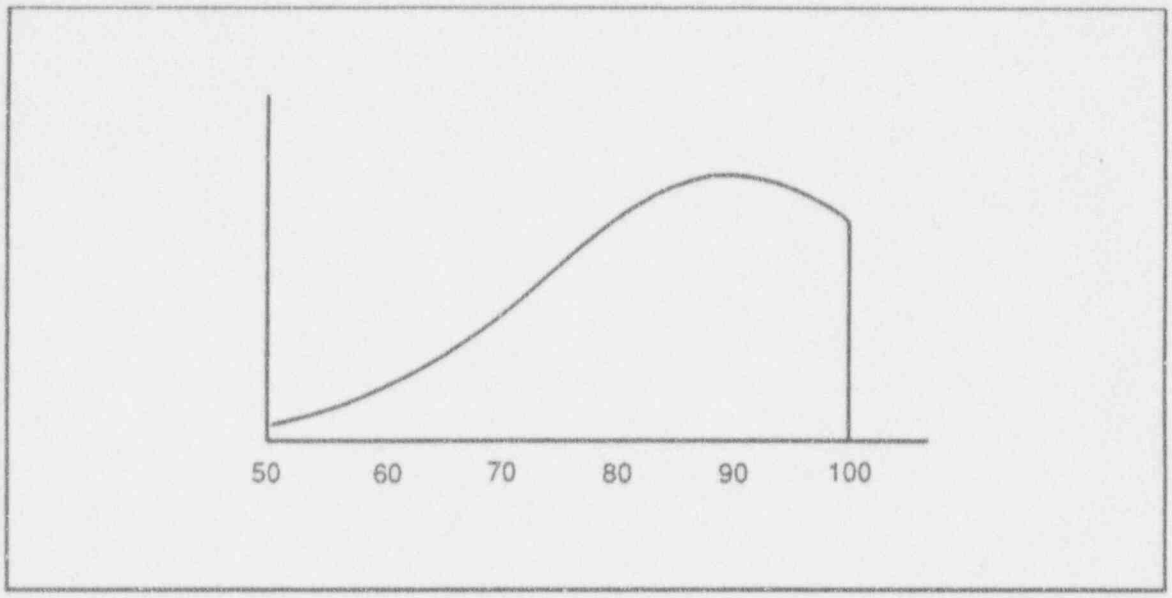



**Figure 8: FREQUENCY POLYGON: A frequency polygon for 30 mechanical maintenance exam scores.**

**b. Curve characteristics**

Four characteristics that define the form of the distribution (curve) are symmetry, skewness, modality, and kurtosis. A curve is symmetrical when the left and right halves of the curve are identical. Many of the non-symmetrical curves are described as skewed. A skewed curve has a "lump" and a "tail." A negatively skewed distribution has a tail on the left, while a positively skewed distribution has a tail on the right. Figure 8 is an example of a negatively skewed distribution. A distribution may have one or more modes, a mode being defined as the most frequently occurring score and appearing as the highest point, or peak, on the curve. A bimodal distribution has two equally high peaks. Kurtosis refers to the relative flatness or peakedness of the curve. Platykurtic distributions are very flat, leptokurtic distributions are peaked, and mesokurtic distributions are in the middle. A single distribution can be symmetric, unimodal, and mesokurtic. A significant curve fitting this description is the normal curve. The normal curve will be discussed after presenting two other concepts: central tendency and variability.

**c. Measures of central tendency**

The central tendency of a distribution can be described in terms of its mode, median, or mean. The mode, as discussed previously, is the most frequently occurring score; it generally occurs near the middle of the distribution but does not have to. There can be more than one mode. The median is the middle score in the distribution. Half of the scores fall above the median, and half fall below the median. For large, relatively normal distributions this is a reliable measure of central tendency. However, a more stable measure for skewed distributions is the arithmetic average, or mean. This value is obtained by summing all the scores and dividing that sum by the number of scores. The mean is the most often used measure of central tendency.

d. **Measures of variability**
Three measures of variability are used. The simplest is the range. Range is defined as the number of score points that the distribution covers. The range is calculated by subtracting the low score from the high score and adding one. Like the mode, the range is a very unstable statistic. Its value is dependent only on the value of the highest and lowest scores in the distribution. The most used measure of variability is known as the standard deviation (SD). The standard deviation is a number that indicates the average amount that the scores in a distribution differ from the mean. If the standard deviation of a distribution is large, the scores are spread over a wide range and the curve is platykurtic. Conversely if the standard deviation is small, the scores are bunched together near the mean, and the curve is leptokurtic. A measure of variability used by many mathematicians is the variance. The variance is merely the square of the standard deviation.

e. **Normal curve**
The normal curve is a bell-shaped distribution that characterizes a large number of naturally occurring phenomenon. The normal curve is unimodal and symmetrical. The mean, median, and mode are all at the same point, the center of the curve. Distances from the mean are often expressed in standard deviation units. Figure 9 shows the "bell curve," standard deviation units, and the percentage of scores that would fall between standard deviations units.

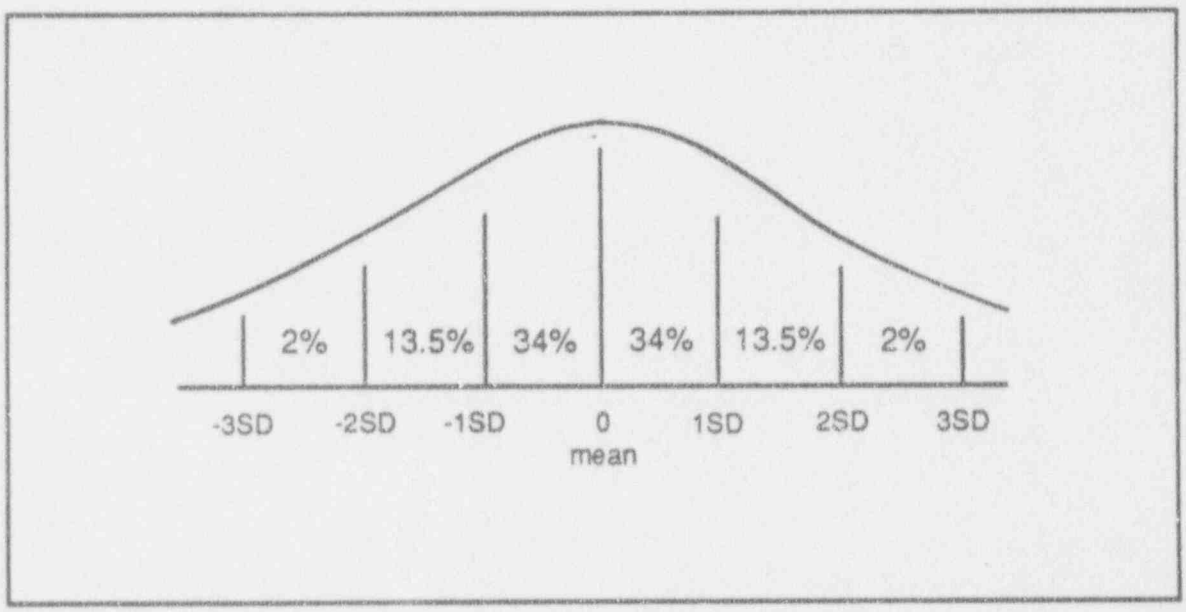



**Figure 9: NORMAL CURVE**

It is useful to know the characteristics of the normal curve since it is the basis for comparison of both single and group scores in a norm-referenced measurement system.

2. **Criterion-referenced interpretation**
Criterion-referenced interpretation poses a different problem than does norm-referenced interpretation. Criterion-referenced interpretation is easy once a criterion level (cutoff score) is established. Any score above the criterion level is acceptable; any score below the criterion

level is unacceptable. The problem, of course, is in determining what the criterion level should be. A cutoff score must be pre-established. Many training programs use 80 percent as a criterion level. The instructor applies this cutoff score to all of the exams he administers. If many trainees fail, the exam is considered hard; if the scores are high, the exam is considered easy. The validity of the 80 percent cutoff score is seldom questioned. A better method is needed to establish criterion levels (cutoff scores).

**a. Establishing the criterion level**

Miller, Williams, and Haladyna have proposed a method for establishing an acceptable level of performance on any examination.[14] The method is defensible and produces consistent results. It is based on a subject matter expert (SME) review of each test item. The subject matter expert may be a single instructor or a team of experts. The ideal would be a team composed of an instructor, an experienced job incumbent, and a supervisor who is also experienced in the job classification being reviewed. Each reviewer would look at each exam item and determine what the acceptable performance level should be for a trainee to be considered competent to enter that job. The reviewer then rates each exam item using a point system. Points are assigned for all items, averaged across raters, and then totaled for the examination. The point total becomes the acceptable level of performance (ALP) for the exam. This value could be any percentage of the total score. It could be 60 percent or it could be 90 percent. Having determined the acceptable level of performance for the criterion-referenced examination, the instructor can be satisfied that regardless of trainee scores, high or low, the examination cutoff score was appropriately established.

Sometimes it is too hard to go against convention. Rather than convince the world that an appropriate ALP may be 60 percent or 90 percent, it may be necessary to mechanically adjust the ALP to a less controversial level, 80 percent for example. This can be done by modifying the difficulty level of the test. By using harder or easier questions the ALP can be adjusted to the desired difficulty level.

---

[14] Miller, H. G., Williams, R. G., and Haladyna, T. M., Beyond Facts: Objective Ways to Measure Thinking, (Englewood Cliffs, N. J.: Educational Technology Publications, 1978).

# GLOSSARY OF TERMS

***Achievement Test:*** An instrument designed to measure a trainee's grasp of some body of knowledge or skill proficiency

***Affective:*** Aspects of an individual that involve emotional feelings rather than intellectual knowledge

***Aptitude Test:*** An instrument designed to assess an individual's potential for performing some task or skill area

***Attitude Inventory:*** An instrument designed to assess an individual's feelings regarding some topic

***Average:*** A score that provides an indication of the typical performance of a group of scores--The mean, median, and mode of a distribution of scores are all commonly used as averages.

***Central Tendency:*** A term referring to the most typical performance of a group of individuals; generally the mean, median, or mode

***Cognitive:*** Aspects of a person that refer to knowledge or understanding

***Completion Item:*** A test item that requires a trainee to supply the missing part of a statement; also referred to as fill-in-the-blank

***Concurrent Validity:*** The degree to which a test provides equivalent results with another test administered at the same time

***Construct Validity:*** The degree to which a test measures a particular theory, concept, or idea

***Content Validity:*** The degree to which a test measures the specific objectives or content of that test

***Correlation Coefficient:*** A numerical value ranging from -1 to +1 that indicates the relationship between two sets of scores or other measures of each individual in a group--A value of 0 indicates no relationship; +1 or -1 indicates a perfect relationship, either positive or negative.

***Criterion:*** A characteristic or combination of characteristics used as the basis for judging a performance

***Criterion-referenced:*** A system of score interpretation where each score is compared to a predetermined standard or cutoff score

***Criterion-referenced Test:*** An examination that uses an established standard or cutoff score as a measure of acceptable performance

***Cutoff Score:*** The score at which a trainee is deemed to have met the criteria on an exam

***Diagnostic Test:*** An instrument that is designed to identify the strengths and weaknesses of an individual for a given content area

***Difficulty Index:*** A numerical index ranging from .00 to 1.00 that indicates the percentage of trainees who answer a test item correctly--An index of .00 indicates that no one answered the test item correctly while an index of 1.00 indicates that all individuals answered the item correctly.

***Discrimination Index:*** A measure of a test item's ability to differentiate between good and poor trainees--A high score indicates that many good trainees and few poor trainees answered the item correctly (good and poor are typically determined by overall test scores but may also be established by an external criteria).

***Distracter:*** An incorrect alternative among the choices of a test item

***Essay Exam:*** A test format that allows trainees to formulate a unique or individual response

***Essay Test Item:*** A question or problem that requires an extensive written response

***Error of Measurement:*** Any difference between an obtained score and a true score on a test is referred to as error of measurement. The actual error of measurement can only be estimated since it is impossible to know what the true score is.

***Equivalent Forms:*** Two or more exams that test the same objectives using different test items or the same test items in a different sequence

***Evaluation:*** The process of judging the quality or worth of an object or activity; the results of that judging process

***Foil:*** An incorrect alternative in a multiple response test item

***Frequency Distribution:*** A graphic display listing scores, or score intervals on one axis of a graph, and the number of trainees at that score or in that interval on the other axis

***Intelligence Quotient (IQ):*** Originally a ratio of an individual's mental age to chronological age, this concept attempts to compare innate abilities between individuals. The "average" person has an IQ of 100 on a scaled score examination.

***Intelligence Test:*** An aptitude test designed to measure an individual's general learning ability (intelligence quotient)

***Item Analysis*** A set of procedures performed on examination items to determine their difficulty and discriminating power

***Item Pool:*** A group of test items covering a defined area--Items for a test can be chosen from this source.

***Item Stem:*** The part of a test item that presents the problem or situation to be solved--The stem may be a question requiring a response or a statement that is followed by the alternatives from which the trainee must choose the best answer.

***Learning Objective:*** A statement of the behavior a trainee is expected to exhibit following instruction

***Matching Item:*** A type of test item in which an individual must recognize associated facts from among many choices

***Mastery Test:*** a term synonymous with criterion-referenced test

***Mean:*** An indication of central tendency--It usually refers to the arithmetic mean, which is computed by summing all the scores of a group and dividing that sum by the number of scores in the group.

***Median:*** A measure of central tendency--The point on a scale of scores that splits the scores in half; 50 percent of the scores are below this point, and 50 percent of the scores are above this point.

***Mode:*** The least reliable of the common measure of central tendency--The mode is the most frequently occurring score in a distribution of scores.

***Multiple Choice Item:*** A test item composed of a stem and several alternatives from which the trainee must select the best answer

***Normal Distribution:*** A theoretical frequency distribution represented by a symmetrical bell-shaped curve; sometimes referred to as the bell curve

***Norm-referenced:*** score interpretation based on the comparison of an individual's score with an appropriate reference group

***Objective Test:*** A test that can be scored without subjective judgment in the scoring

***Performance Test:*** A test that requires the trainee to demonstrate skill by actual operation or manipulation of tools and equipment

***Predictive Validity Evidence:*** The ability of a test to forecast future performance on a subsequent measure

***Psychomotor:*** The domain of human performance that relates to physical performance based on mental activity

***Range:*** The smallest interval on a scale of scores that will include all scores, mathematically defined as the largest score minus the smallest score plus one

***Raw Score:*** The numerical score first assigned when scoring a test before conversion to a derived score

***Reliability:*** The consistency or repeatability of any measure

***Score:*** A numerical indication of the performance an individual displays on a test

***Short Answer Test Item:*** A form of question that requires a limited response, usually a word, phrase, or number

***Split-Halves Reliability Coefficient:*** A correlation coefficient created from an examination by considering one half of an exam as a separate test from the other half--The correlation between these two halves provides an estimate of the reliability of the total test.

***Standard Deviation:*** A measure of variability of a set of scores around the group mean--The SD is mathematically defined as the square root of the mean of the squared deviations of the scores from the mean of the distribution.

***Standard Error of Measurement:*** An estimate of the standard deviation of the errors of measurement associated with the test scores in a given test

***Standardized Test:*** A test that has the directions, time limits, and conditions of administration made consistent for all offerings of the test

***Statistic:*** A numerical value computed on a sample of data

***Test:*** A measurement instrument; examination

***True/False Item:*** The simplest form of a multiple choice item containing only two alternatives, true or false

***True Score:*** The ideal or correct score for an individual--Its value cannot be known, but it can be estimated when assumptions regarding error of measurement are made.

***Validity:*** The degree to which a test measures what it purports to measure

***Variance:*** A measure of the spread of individual scores about the group mean, equivalent to $SD^2$

# REFERENCES


Anastasi, A. Psychological Testing, 3rd ed. Toronto, Ontario: Macmillan Company, 1968.

Beggs, D.I., and Lewis, E.L. Measurement and Evaluation in the Schools. Boston, MA.: Houghton Mifflin Company, 1975.

Bloom, B.S., Englehart, M.D., Furst, E.J., Hill, W.H., & Krathwohl, D.R. Taxonomy of Educational Objectives. Handbook I: The Cognitive Domain. New York, N.Y.: David McKay Company, 1956.

Cronbach, L.J., "Coefficient Alpha and the Internal Structure of Tests," Psychometrika, Vol. 16, pages 297-334, 1951.

Ebel, R.L., and Frisbie, D.A. Essentials of Educational Measurement, 4th ed. Englewood Cliffs, N.J.: Prentice-Hall, 1986.

Erickson, R.C. and Wentling, T.L. Measuring Student Growth: Techniques and Procedures for Occupational Education. Boston, MA.: Allyn and Bacon, Inc., 1976.

Glass, G.V. and Stanley, J.C. Statistical Methods in Education and Psychology. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1970.

Green, B.F., "A Primer of Testing," American Psychologist, Vol. 36, No. 10, pages 1001-1011, American Psychological Association, Inc.,1981.

Green, J.A. Teacher-Made Tests, 2nd ed. New York, N.Y.: Harper and Row Publishers, Inc., 1975.

Gronlund, N.E. Measurement and Evaluation in Teaching. New York, N.Y.: Macmillan, 1971.

Hills, J.R. Measurement and Evaluation in the Classroom. Columbus, OH.: Charles E. Merrill Publishing Co., 1976.

Kane, M.T., and Brennan, R.L., "Agreement Coefficients as Indices of Dependability for Domain-referenced Tests," Applied Psychological Measurement, Vol. 4, pages 105-26, 1980.

King, W.A. and Lamberg, W. Criterion-referenced and Norm-referenced Tests: A Comparison. Ann Arbor, MI.: Ulrich's Books, Inc., 1964.

Krathwohl, D.R., Bloom, B.S., & Masia, B.B. Taxonomy of Educational Objectives, Handbook II: The Affective Doman. New York: David McKay Co., 1964.

Mehrens, W.A., and Lehmann, I.J. Measurement and Evaluation in Education and Psychology, 3rd ed. New York, N.Y.: Holt, Rinehart and Winston, 1984.

Miller, H.G., Williams, R.G., and Haladyna, T.M. Beyond Facts: Objective Ways to Measure Thinking. Englewood Cliffs, N.J.: Educational Technology Publications, 1978.

Noll, V.H., Scallell, D.P., and Craig, R.C. Introduction to Educational Measurement, 4th ed. Boston, MA.: Houghton Mifflin Co., 1979.

Rahmlow, H.F. and Woodley, K.K. Objectives-based Testing: A Guide to Effective Test Development. Englewood Cliffs, N.J.: Educational Technology Publications, 1979.

Roid, G.H. and Haladyna, T.M. A Technology for Test item Writing. Orlando, FL.: Academic Press, Inc., 1982.

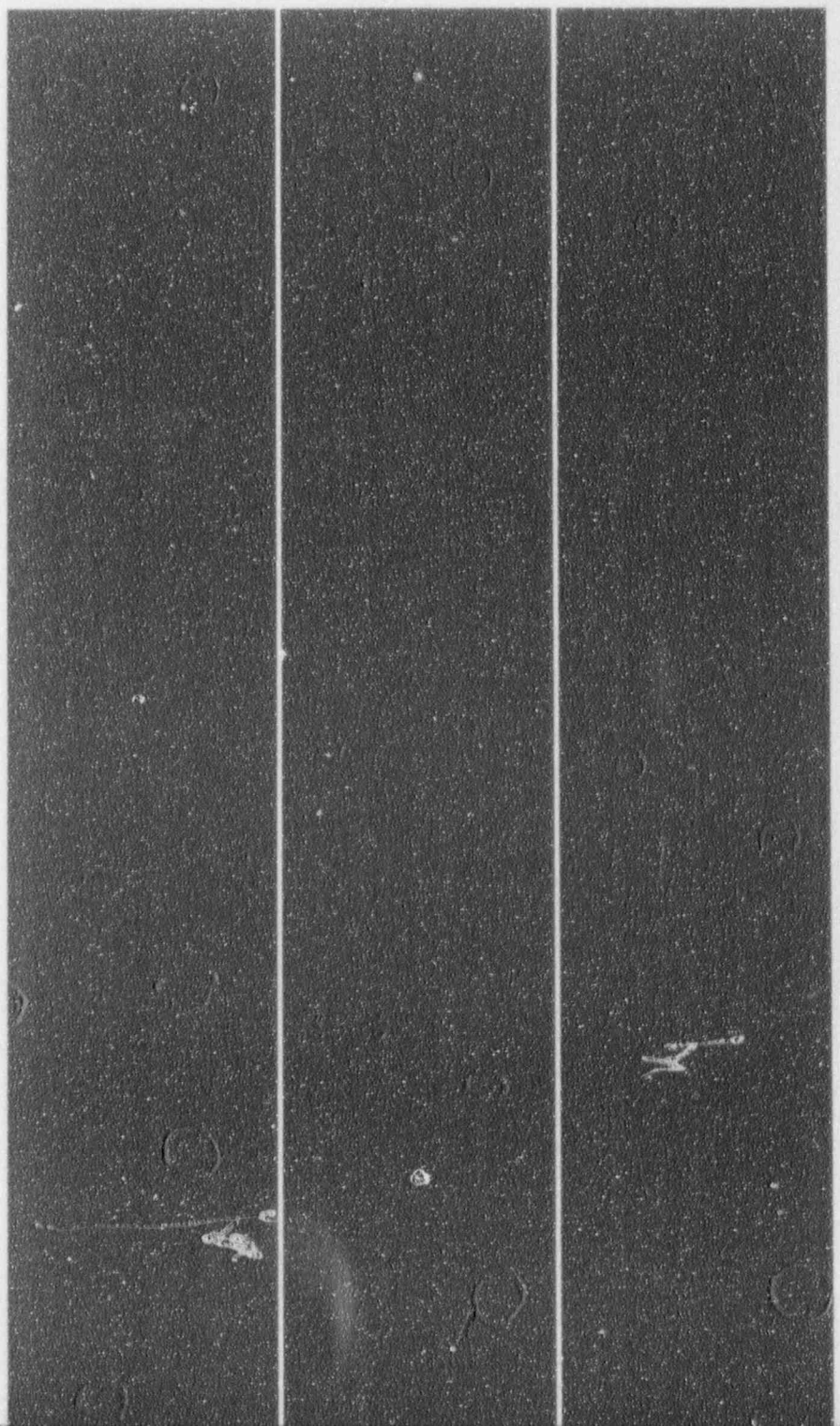
Simpson, E.J., "The Classification of Educational Objectives: Psychomotor Domain," *Illinois Teacher of Home Economics*, 10(4), pages 110-144, 1966.

(NPC) is partially supported by assistance from the Tennessee Valley Authority (TVA), a Federal agency. Under Title VI of the Civil Rights Act of 1964 and applicable TVA regulations, no person shall, on the grounds of race, color, or national origin, be excluded from participation in, be denied the benefits of, or be otherwise subjected to discrimination under this program. If you feel you have been excluded from participation in, denied the benefits of, or otherwise subjected to discrimination under this program on the grounds of race, color, or national origin, you or your representative, have the right to file a written complaint with TVA not later than 90 days from the day of the alleged discrimination. The complaint should be sent to Tennessee Valley Authority, Office of Equal Employment Opportunity, 400 Commerce Avenue, EPB 14, Knoxville, Tennessee 37902. The applicable TVA regulations appear in Part 1302 of Title 18 of the Code of Federal Regulations. A copy of the regulations may be obtained on request by writing TVA at the address given above.

Printed in U.S.A.

GENERAL DISTRIBUTION

September 1993
ACAD 88-002 (Addendum II)
ACADEMY DOCUMENT



NATIONAL
ACADEMY
FOR NUCLEAR
TRAINING

700 Galleria Parkway
Atlanta, Georgia 30339-5957
Telephone 404-644-8543

UNITED STATES
NUCLEAR REGULATORY COMMISSION
WASHINGTON, D. C. 20555

April 19, 1994

MEMORANDUM FOR: Darlene Huyer
Anstec, Inc.

FROM: Tremaine Donnell, INPO Coordinator
Records and Archives Services Section
Information and Records Management Branch
Office of Information Resources Management

SUBJECT: ESTABLISHMENT OF DATA RECORD FOR INPO DOCUMENTS

The Records and Archives Services Section has received the attached INPO Document.

Distribution Code: NYF2

Comments: This is a **General Distribution Document**, copyrighted by INPO. The Institute authorizes the NRC to place this document in the Public Document Room. The document is covered within the Copyright License executed between the NRC and INPO on December 8, 1993.

Please return RIDS distribution to Tremaine Donnell, 5C3, Two White Flint North, 415-5633.

Tremaine Donnell

Tremaine Donnell, INPO Coordinator
Records and Archives Services Section
Information and Records Management Branch
Office of Information Resources Management

Enclosure: As stated

PLEASE NOTE: Hard copy is available from the NRC File Center.

cc: JDorsey