

**BY ORDER OF THE  
SECRETARY OF THE AIR FORCE**

**AIR FORCE HANDBOOK 36-2235  
VOLUME 12**

**1 NOVEMBER 2002**

***Personnel***



**INFORMATION FOR DESIGNERS OF INSTRUCTIONAL SYSTEMS  
TEST AND MEASUREMENT HANDBOOK**

---

**RECORDS DISPOSITION:** Ensure that all records created by this handbook are maintained and disposed of IAW AFMAN 37-139, "Records Disposition Schedule"

---

**NOTICE:** This publication is available digitally on the AFDPO www site at: <http://afpubs.hq.af.mil>.

---

OPR: HQ AETC/DOZ (Gary J. Twogood)  
Supersedes: AFH 36-2235, Volume 12, 1 February 1998

Certified by: HQ USAF/DPDT  
(Col Patricia L. C. Priest)  
Pages: 240/Distribution: F

---

This publication requires collecting and maintaining information protected by the Privacy Act of 1974 authorized by 10 USC 8013. Privacy Act system notices F036 AF PC C, Military Personnel Records System; and OPM/Govt 1, General Personnel Records apply.

This volume provides information and guidelines for applying current instructional test and measurement learning theory, described in AFMAN 36-2234 and AFH 36-2235 Volumes 5 and 7, to the design and development of Criterion-Referenced tests and the measurement of student performance, and information and guidelines for conducting validation of instructional resources and evaluation of instructional systems. This handbook is a review of current Air Force test and measurement learning theory; and a guide for Air Force personnel who develop Criterion-Referenced test and measurement instruments, and who validate instructional resources and operationally evaluate instructional systems.

<b>Chapter 1 OVERVIEW OF USAF LEARNING THEORY FOR TEST AND MEASUREMENT .....</b>	<b>5</b>
Section A Test Development.....	6
Section B Computer-Managed Instruction (CMI) and Interactive Courseware (ICW) Test and Measurement .....	15
Section C Performance Proficiency Test and Measurement.....	19
Section D Types of Intellectual Skills .....	23
<b>Chapter 2 GUIDELINES FOR CRITERION-REFERENCED TEST AND MEASUREMENT DEVELOPMENT AND USE.....</b>	<b>28</b>
Section A Introduction.....	29
Section B Test Validity .....	31
Section C Test Reliability .....	33
Section D Retention and Transfer Tests .....	37
Section E Overview of Predictive Written Tests .....	46
Section F Common Types of Predictive Written Test Items.....	48
Section G Validity of Predictive Written Tests .....	51
Section H Performance Test and Measurement .....	54
Section I Types of Process Rating Methods and Ways to Avoid Rating Errors.....	60
<b>Chapter 3 GUIDELINES FOR CONSTRUCTING CRITERION-REFERENCED TESTS AND USING THE SURVEY TEST .....</b>	<b>70</b>
Section A Introduction.....	71
Section B Guidelines for Constructing Criterion-referenced Tests .....	73
Figure 1 Notional Test Construction Worksheet for Product Test Items.....	87
Figure 2 Notional Test Construction Worksheet for Performance Test Items.....	88
Section C Guidelines for Using the Survey Test .....	89
<b>Chapter 4 GUIDELINES FOR VALIDATION OF INSTRUCTIONAL RESOURCES .....</b>	<b>93</b>
Section A Introduction.....	94
Section B Validation of Resource Requirements .....	95
Section C Validation of the Instructor.....	100
Section D Validation of the Instructional Materials .....	105
Section E Analyzing Tryout Test and Measurement Items and Making Revisions to Instructional Materials.....	111
Section F Conducting a Validation of the Instructional System .....	125
<b>Chapter 5 GUIDELINES FOR OPERATIONAL EVALUATION OF INSTRUCTIONAL RESOURCES .....</b>	<b>132</b>
Section A Introduction.....	133
Section B Internal Evaluation .....	136
Section C External (Field) Evaluation.....	144
Section D Questionnaires .....	146
Section E Field Visits .....	153
Section F Job Performance Evaluation .....	157
<b>Chapter 6 ADVANCED MEASUREMENT TOPICS .....</b>	<b>162</b>

Section A Guidelines for Standard-Setting for Norm-Referenced and Criterion-referenced Tests..... 164

Section B Criterion-Referenced Test Analysis ..... 183

Section C Guidelines For Test Validation..... 195

Section D Lessons Learned for Test Item Development..... 202

Section E Summary of Lessons Learned for Test Item Development..... 218

**Attachment 1 – GLOSSARY OF REFERENCES AND SUPPORTING INFORMATION..... 229**

## Chapter 1

# OVERVIEW OF USAF LEARNING THEORY FOR TEST AND MEASUREMENT

---

**Purpose of this chapter**

The information in this chapter is to be used in conjunction with the information contained in AFM 36-2234, Instructional System Development, and in AFH 36-2235, Information for Designers of Instructional Systems, Volumes 1-11.

The purposes of this chapter are to:

- Provide an overview of current learning theory for test and measurement.
  - Describe test and measurement theory for Computer-Managed Instruction (CMI) and Interactive Courseware (ICW).
  - Describe performance proficiency test and measurement.
  - Define the types of intellectual and motor skills.
- 

**Where to read about it**

This chapter contains four sections:

<b>Section</b>	<b>Title</b>	<b>Page</b>
A	Test Development	6
B	Computer-Managed Instruction	15
C	Performance Proficiency Test and Measurement	19
D	Types of Intellectual Skills	23

---

## Section A Test Development

---

### Introduction to test development

Test development has three major requirements:

Good tests adequately measure the instructional objectives they support.

The performance required in the test should match the performance required in the objective.

Tests should be prepared immediately after the objective is written.

---

### Purpose of testing

The primary purpose of testing is to assess student attainment of the behaviors specified in the terminal and enabling objectives.

---

### Secondary purposes of testing

Tests also serve several secondary purposes such as:

Identifying problems or weaknesses in the instruction.

Indicating whether a class is performing up to standards on specific objectives.

Indicating the capability of the instructor and the instructional medium to facilitate learning.

---

### Types of tests

The basic types of tests used in the Air Force are:

Type of Test	Purpose of Test
<b>Criterion</b>	<p>Used to measure the student's attainment of an objective.</p> <p>Used to measure the effectiveness of the instruction.</p> <p>Used to measure the student's ability to attain each objective.</p>
<b>Pre-test</b>	<p><b>Readiness Pre-test</b></p> <p>Used to measure prerequisite course entry skills.</p>

---

**Types of tests  
(Continued)**

<b>Type of Test</b>	<b>Purpose of Test</b>
<b>Pre-test (Continued)</b>	<p><b>Placement Pre-test (Adaptive Pre-test)</b>            Used to measure attainment of course or unit objectives.            Used after the instructional system becomes operational to determine how much instruction individual students need. (What terminal or enabling objectives were not previously mastered?)</p> <p><b>Diagnostic Pre-test</b>            Used to determine attainment of supporting intellectual and motor skills (enabling objectives) necessary to master a terminal objective.            Used during validation of instruction to measure the effectiveness of instruction, and to identify and correct weaknesses in the instruction.            Contain a number of test items in each specific subject area to allow a detailed search for a source of learning deficiencies, what the student needs to learn, etc.</p> <p><b>Survey Pre-test</b>            Used to determine what prospective students already know and can do before receiving instruction.            Used during development of instruction to gather data for design of instruction.</p>
<b>Post-test</b>	<p>Used after exposure to an instructional program to provide a measure of the changes that have occurred during instruction.            Used to compare the capabilities of an individual to those of other students (grading on the curve).</p>
<b>Norm-Referenced</b>	<p>Not appropriate for Criterion-Referenced instruction.            Used for Air Force selection tests.</p>

**Test construction  
for types of  
learning**

The outcomes of planned instruction consist of student performances, which show that capabilities have been acquired. The types of learning are *intellectual skills, verbal information, cognitive strategies, motor skills, and attitudes*. (The types of learning are further described in Section D of this chapter.)

Assess student performance to determine whether the newly designed instruction has met its design objectives.  
Conduct assessment to learn whether each student has achieved the set of capabilities defined by the instructional objectives.

Type of Learning Outcome	Best Method of Testing	Activities That Indicate Achievement of Objectives
<b>Intellectual Skills</b>	<b>Predictive Tests</b>	
Discriminations	Multiple-choice and true/false	Detect similarities or differences.
Concrete Concepts / Defined Concepts	Constructed response (labeling, sorting, matching)	Recognize examples or non-examples.
Rule Learning	Performance of integrated tasks or constructed response(short answer)	Apply rule, principle, or procedure. Solve problems. Produce a product.
Verbal Information	Performance Tests Constructed response (fill in the blank, essay questions, oral testing)	State information verbally or in writing.
Cognitive Strategies	Performance Tests Student explains process to test administrator. (Oral testing)	Self-report or audit trail of work done. State strategies and tactics, and expected results of actions.

**Test construction  
for types of  
learning  
(Continued)**

Type of Learning Outcome	Best Method of Testing	Activities That Indicate Achievement of Objectives
Motor Skills	Performance Tests	Perform smooth, timely coordinated action.
Attitudes	Performance Tests Observe student in different situations.	Display desired situated behavior.

**Test  
characteristics**

Consider these characteristics when developing tests to ensure that the tests measure what is intended each time they are administered:

Characteristic	Definition
Validity	Content Degree to which the test measures what it is intended to measure. Predictive Degree to which the test predicts performance.
Reliability	Degree to which the test consistently yields the same results. Test-Retest Reliability Consistency across two administrations to the same students. Split-Halves Reliability Consistency across two forms of the same test.
Usability	Tests that are easy to administer, score, and interpret.

**Assessment  
methods**

Most Air Force tests can be classified into two main groups: *predictive test and performance tests.*

**Performance tests**

*Performance test* is one in which the student actually *performs the skill required by a terminal or enabling objective.*

---

**Written or verbal performance tests**

Points to consider are:

*Discrimination, concrete concept, and defined concept* intellectual skills may be tested by writing them on a piece of paper, entering them into a computer, or stating them orally. These are examples of *written or verbal performance tests*. *Rule-learning* and *verbal information* intellectual skills may also be tested by using written or verbal performance tests. *Cognitive strategy* intellectual skills may be tested by using *verbal performance tests*.

---

**Psychomotor performance tests**

Many types of tasks, especially equipment operation tests, involve many different intellectual and motor skills that have to be performed in an integrated manner.

Combined intellectual skills and motor skills *associated with performance of a hands-on task* are called *psychomotor skills*. A test that measures combined intellectual and motor skills associated with a hands-on task is called a *psychomotor performance test*.

Example, the *psychomotor task* of bleeding a hydraulic brake system involves

- Recall of a procedure (rule learning intellectual skills).

- The physical performance of the steps (motor skills).

- Recognition of the parts and tools (discrete concepts intellectual skills).

- Cleanliness and safety (attitude skills).

---

**Motor skill performance tests**

Motor skill performance cannot be measured with a written or oral test. Motor skill performance tests:

- Require a real or operational mock-up of equipment or computer-generated simulations of equipment operation.

- Require the student to demonstrate mastery of an actual operational hands-on task.

- Have content validity. *The most content-valid test of any kind of learning is an operational hands-on task performance test.*

- Are generally time-consuming because they often have to be conducted one-on-one with real equipment or simulators.

---

---

**Predictive tests**

If the actual operational behavior required for job performance (terminal learning objective) cannot be tested in the instructional system using a performance test, the next best option is to test behaviors that enable performance of the desired skill (enabling learning objectives), and from that information make a prediction as to whether the student would be able to perform the operational task.

For example, if a student could write the steps for bleeding a brake system, there is a better probability that the student could actually perform the task than someone who did not know the steps.

Predictive tests *are valid to the extent that they predict student performance on the actual task.*

---

**Types of predictive tests**

The most common types of predictive written test questions are essay, short answer, fill-in-the-blank, labeling, multiple-choice, matching, and true-false.

Computer-based predictive tests can use different types of input systems that have a high degree of fidelity with real-world tasks.

A simple input device such as a joystick or mouse allows for identification by pointing with a cursor.

More elaborate devices such as magnetic field detectors, infrared detectors, head-eye tracking devices, etc., allow the computer to detect even more complex behavior.

---

**Comparison of performance and predictive test items**

The best type of test is one that provides accurate information regarding the student's mastery of the objective.

Different types of test items have to be considered in terms of validity and reliability of the test.

The differences between predictive and performance test items are:

---

**Comparison of performance and predictive test items (Continued)**

Predictive Test Item	Performance Test Item
<p>Requires students to demonstrate mastery of <b>enabling objectives</b> by responding to various types of written, oral, or computer-generated questions. Emphasizes intellectual skills related to a performance objective. May require students to find, read, and use technical materials. Items are intellectual skills that require mastery to enable job performance. Items are independent questions, and the test item sequence will not always affect the outcome of the test. Errors on one test item do not always affect performance on another item.</p>	<p>Requires students to demonstrate mastery of <b>terminal or enabling objectives</b> by responding to various types of written, oral, or computer-generated questions or by performing a job task under controlled conditions. Emphasizes intellectual skills associated with the hands-on performance of a motor skill (psychomotor skills). May require students to find, read, and use certain technical materials. (job aids, for example). Items are often sequential intellectual or motor skills. Errors early in the performance sequence will often affect the final outcome of the task.</p>

**Test construction factors**

Several key factors should be considered when constructing tests to determine what to measure, the level of testing, learning levels, test length, and the selection and arrangement of test items.

**What to measure**

Guidelines for *what to measure*:

Perform an analysis of the terminal and enabling objectives to identify what intellectual skills and motor skills should be measured.

List the tasks to be performed and the terminal and enabling objective behaviors to be covered by the test.

One or more test items are required to adequately measure each terminal and enabling objective behavior.

Design tests to measure all of the intellectual and motor skills required to master each enabling and terminal objective behavior.

<b>Levels of testing</b>	Guidelines for the <i>level of testing</i> :  Correlate the level of testing with the level of learning for each enabling and terminal objective behavior.
<b>Bloom's learning levels</b>	Bloom's Taxonomy of <i>levels of learning</i> for the cognitive domain are:  <i>Knowledge</i> <i>Comprehension</i> <i>Application</i> <i>Analysis</i> <i>Synthesis</i> <i>Evaluation</i>
<b>Air Force levels of learning</b>	AFM 36-2234 defines the levels of learning for intellectual skills as:  <i>Discriminations</i> <i>Concrete Concepts</i> <i>Defined Concepts</i> <i>Rule Learning</i> <i>Verbal Information</i> <i>Cognitive Strategies</i>
<b>Test length</b>	Guidelines for <i>test length</i>  Ensure adequate coverage of each terminal and enabling objective. Longer tests are normally more reliable.
<b>Selection of test items</b>	Guidelines for <i>selection and arrangement of test items</i> :  Select test items that cover the most essential and significant teaching points. Select test items that are clear, concise, and well written to minimize misunderstandings. Group items of the same type together, if possible. Arrange individual test items in approximate order of difficulty.

## **Section B**

### **Computer-Managed Instruction (CMI) and Interactive Courseware (ICW) Test and Measurement**

---

**Introduction**

An important aspect of ICW development is test and test item design, and the design of computer-managed instruction functions and records.

---

**Definition of CMI**

CMI is the function of the ICW authoring software related to student test and measurement data collection.

---

**CMI functions**

CMI generally includes the following functions:

**Administrative**

Registration of the student in an ICW course.

Point-of-entry for the student into the course, often based on a pre-test performance or previously “bookmarked” location. Students should be able to leave a lesson and return to the same point at a later time.

Documentation of the student path through the ICW, and the time spent on specific lessons, segments, or topics.

Disenrollment of students from the course.

**Performance Tracking**

Employment of different types of test items (e.g., digitized video, graphic and animated images).

Collection of data regarding the student’s performance on tests and practice exercises.

Use of error performance and time performance metrics for test items and test segments.

Provision of immediate feedback to the student for test questions on the pre-test, embedded tests, lesson or segment tests and post-tests.

Determination of student mastery of objectives.

Reporting of student performance information.

---

**CMI capabilities**

Prior to designing the CMI for an ICW course, review the selected authoring software to determine the extent of data collection and analysis that is possible.

---

---

**Design of ICW tests**

Develop ICW tests to measure the intellectual skills related to and associated with each hands-on task or terminal instructional objective.

---

**Types of ICW tests**

The types of tests usually developed in ICW courses are pre-tests and criterion tests.

<b>Pre-test</b>	Use a pre-test to measure the student's ability to attain each objective before developing ICW and before entering students in an ICW lesson.
<b>Criterion Test</b>	Use a criterion test to measure the student's attainment of the objectives and to measure the effectiveness of the ICW.

---

**Guidelines for designing ICW tests**

Table 1 provides guidelines for designing ICW tests.

Table 1 Guidelines for Developing ICW Tests

#	Guideline Description	Rationale
1	Use a student's pre-test score to branch the student to "need to know" information.	Reduces student boredom by not forcing them to review things they already know.
2	Use the student's pre-test score to gauge deficiencies in entry-level prerequisite skills and knowledge.	Stimulates recall of relevant prior knowledge (one of the "events" of instruction).
3	For pre-tests, explain that students are not expected to know all the answers.	Puts students "at ease" with the instruction.
4	Introduce the test by telling students how many questions they will see and how long it should take them to complete the test.	Helps students gauge how extensive the test is.
5	Let students "back out" of taking a pre-test if they know they do not know the content.	Forcing students to take a test when they know they don't know the content can introduce unnecessary stress into a learning situation.
6	Provide clear instructions for taking the test, including how to change answers.	Reduces the possibility of students making errors when they actually have mastered the objective.
7	Provide a method for students to review their completed test.	If students responded with a wrong answer and subsequently realize it, they should be able to correct the answer, just as they can in a paper-and-pencil testing situation.
8	Provide immediate feedback to student answers in the same order that they answered the questions.	Reduces confusion and increases the learning value of a test.
9	Design the program such that the computer "works through" a problem (provides real-time help) interactively for students instead of just giving the correct answer.	Reduces learning time because a student may have a partially correct answer. The computer should identify the point where the student is in error and invite the student to go on from there.
10	If questions are drawn from a "pool of questions," remove correctly answered questions from the pool for subsequent iterations of test items to the student.	Learning criterion has been achieved and students should not be required to answer these questions again.

---

**Additional  
information**

For additional information on determining CMI and testing strategies, see:

Kulhavy, R.W. (1977). Feedback in Written Instruction. *Review of Educational Research*. 47(1), 211-232.

Rattanapain, V. (1992). *The Effects of Learner Characteristics and an Evaluation Override Option on Achievement, Attitude, and Pattern of Program Use in Computerized Drill and Practice*. Ph.D. thesis, The Pennsylvania State University.

Shapiro, A.F., and Gibbs, W.J. (1993). *Design Considerations When Building Multimedia Instructional Systems*. Multimedia and Videodisc Monitor. March, 17-21.

Tennyson, R.D. (1980). Instructional Control Strategies and Content Structure as Design Variables in Concept Acquisition Using Computer-Based Instruction. *Journal of Educational Psychology*. 72(4), 525-532.

Wager, W., and Wager, S. (1985). Presenting Questions, Processing Responses, and Providing Feedback in CAI. *Journal of Instructional Development*. 8(4), 2-8.

---

## Section C

### Performance Proficiency Test and Measurement

---

**Method for hands-on performance assessment**

Assess student hands-on performance proficiency for each training objective by using the following rating scale:

The scale begins with total instructor demonstration (level 1.0), and ends with no instructor intervention (level 4.0). *Initial proficiency for a psychomotor skill* is defined by achievement of level 2.5 (the mean level).

<b>Grading Criteria For Hands-on Performance Assessment</b>	
Level 1.0	The student demonstrated a lack of knowledge about the task or made major deviations or omissions that made accomplishment of the task impossible. The instructor was required to demonstrate proper accomplishment of the task.
Level 1.5	The student demonstrated limited knowledge of the task. Although the student can begin the task, performance deteriorates quickly and extensive instructor interaction is required to maintain safe accomplishment.
Level 2.0	The student has a basic understanding of the task, but errors or deviations are significant and would jeopardize safety or mission accomplishment. Even under ideal conditions, extensive instructor intervention is required for safety or mission accomplishment.
Level 2.5	The student made errors or deviations. Limited assistance along with frequent coaching by the instructor was essential for safe accomplishment of the task. The student has sufficient systems knowledge to make correct responses when provided coaching by the instructor.
Level 3.0	The student accomplished the task successfully, but there were slight errors or deviations that the student could not correct. The instructor was required to provide coaching for smooth performance, but not for safe mission accomplishment. The student can perform under ideal conditions, but would have difficulty under adverse conditions.

**Method for hands-on performance assessment (Continued)**

<b>Grading Criteria For Hands-on Performance Assessment (Continued)</b>	
Level 3.5	The student was able to accomplish the task safely and successfully with minor errors or deviations. The student was able to correct these minor errors and no assistance was required from the instructor.
Level 4.0	The student performed the task without errors or deviations. No instructor intervention was required. The student has progressed beyond mere proficiency and could probably perform well under adverse conditions.

**Using the grading scale**

This scale has been used successfully as grading criteria on student progress in both simulation and inflight aircrew training. The scale is the basis for rating forms that track student performance on each training event.

**Progress report form**

Give the student a rating for each training event that is accomplished during a training session.

Use a Progress Report Form with the rating scale across the top, and the training events listed down the side.

Provide space to the side of each event for recording the rating and any instructor notes.

**Guidelines for assessment of progress and difficulties**

Record student actions as a basis for later analysis or diagnosis of difficulties.

Record the type and frequency of errors made during practice. Maintain separate records for each student.

Maintain cumulative records across students for assessing the training system.

**Diagnosis of difficulties**

Points to consider are:

Slow student progress may indicate a deficiency in the development of their cognitive strategies or metaskills and/or lower-level intellectual skills.

---

**Diagnosis of difficulties (Continued)**

Substantial skills degradation after relatively brief periods of no practice indicate that cognitive strategies or metaskills and/or lower-level intellectual skills may be lacking. Remedial training may get a student through a performance check, but intellectual skill deficiencies will affect job performance.

---

**Specific reasons for lack of student progress**

If student progress seems to slow down or stop too soon, check for:

Lack of adequate perception and encoding of cues associated with on-task performance (poor short-term retention and judgment ability).

Lack of recall of details of what they have just done and why (poor judgment and decision-making ability).

Persistence in the same inadequate actions (misunderstanding of the problem, and/or lack of meaningful checkpoints for self-assessment).

Inability to adapt to changes in task requirements or conditions (inadequate cue perception and encoding, and the generation of appropriate actions).

Tendency to simply react to cues and task conditions with no plans or even knowledge of what to do or expect (no expectancies; failure to stay ahead of system).

Poor retention of basic task-situational requirements (lack of organization of task knowledge).

Tendency to react in ways that have undesirable outcomes (failure to assess effects of actions).

Undue delays in student responses (confusion regarding cues, their interpretations, action requirements; cue and/or action interference; insufficient mastery of skill components).

Need for excessive guidance (inadequate skill understanding, or inability to make effective use of knowledge).

---

---

**Student failures contributing to lack of progress**

Student failures can contribute to the lack of progress:

Failure to recognize feedback (poor perception and encoding of cues).

Failure to note, assess, and prioritize all cues (cue perception deficiency).

Failure to discriminate false or irrelevant cues (cue encoding deficiency).

Failure to select proper actions (generation of cognitive strategy deficiency).

---

**Corrective action for student plateaus**

Students will often reach temporary “plateaus” where, to a casual observer, progress has ceased. It is at such points that significant new integration of skill components normally occur, resulting eventually in a fairly rapid increase in proficiency.

If a plateau seems to occur too soon (i.e., at too low a level of proficiency), or last too long, there is a need to find out why and correct the underlying difficulty.

If a plateau seems to persist too long, skill integration is likely the problem.

Provide students with non-threatening updates regarding their progress.

Be specific as to successes and shortcomings.

---

## Section D

### Types of Intellectual Skills

---

**Overview**

Intellectual skills are the foundation for all higher learning.

Lower-level intellectual skills include associations, discriminations, concrete and defined concepts, rule-using, and verbal information.

Higher-level intellectual skills are cognitive strategies (generation of strategies and tactics) and metacognition (verbalization of judgment and decision-making schema). Develop predictive or performance test items to measure the intellectual skills associated with all terminal and enabling objectives in a course of instruction.

---

**Hierarchical nature**

Intellectual skills are hierarchical in nature.

In order to learn higher-order intellectual skills, the learner should possess the prerequisite lower-level intellectual skills. For example, to learn a rule or principle, the student must understand the prerequisite associations, discriminations, concrete concepts, and discrete concepts as well as the relationships among the concepts.

---

**Critical judgment and decision-making skills**

Critical judgment and decision-making components of intellectual skills must be formally taught and evaluated. Critical judgment and decision-making components of intellectual skills are highlighted in the following descriptions.

---

**Associations and discriminations**

*Associations* are skills related to knowing the names or characteristics of the cues associated with a physical object or concept. Knowing what critical cues to perceive.

*Discriminations* are skills related to the *perception of cues*. Knowing what characteristics of critical cues to perceive.

---

**Concrete concepts (classifications)**

Skills related to *classifying physical objects* into one or more classes based on their *physical attributes*. Encoding or classifying perceived cues as a critical condition.

---

---

<b>Defined concepts (classifications)</b>	Skills related to <i>classifying symbolic objects</i> into one or more classes based on a <i>definition</i> . The definition is a <i>rule for classification</i> . Encoding or classifying the probable cause for a critical condition cue.
<b>Rule learning (rule-using and problem-solving)</b>	Skills related to <i>applying documented or job-aided principles or procedures to solve problems</i> . Problem-solving is the ability to <i>discriminate and classify job performance conditions, to recall relevant principles or rules, and to use them to solve a problem</i> . The product of problem solving is not only the solution to the problem, but may involve the <i>discovery of a new rule or procedure (lessons learned)</i> to be used in a similar situation. Generating the appropriate rule (strategy) and steps (tactics) to perform for documented or job-aided procedural segments that require memorization.
<b>Verbal information</b>	(Verbalizing Lower-Level Intellectual Skills) Skills related to stating learned associations, discriminations, concrete concepts, defined concepts, and learned rules for documented or job-aided procedural segments that require memorization (e.g., Perform Normal Takeoff, Engine Fire on Takeoff).
<b>Cognitive strategies and tactics</b>	(Verbalizing Strategies and Tactics) Skills related to the ability of an individual to state the appropriate rule (strategy), and steps (tactics) to perform in response to non-documented or job-aided critical conditions (e.g., Excessive EGT Indications on Takeoff).
<b>Metacognition</b>	(Verbalizing Judgment and Decision-Making Schema) The ability to state the totality of the non-documented or job-aided lessons learned, normal condition results expected, normal condition steps (tactics), critical cues, tactics in response to critical cues, and tactics in response to critical condition cues upon termination of a procedural segment (e.g., Judgment and Decision-making Schema for Perform Engine Start).
<b>Motor skills</b>	Motor skills are learned behaviors that involve the smooth coordinated use of muscles.

---

---

**Test and measurement of motor skills**

Motor skills most often involve a sequence of activities that may be described verbally as a procedural set of sequential actions. To measure motor skills:

Use *predictive oral tests* to predict performance of motor skills. All motor skills have associated intellectual skills. These associated intellectual skills are *enabling behaviors for performance* of the motor skill. Enabling intellectual skills associated with the performance of motor skills should be tested by predictive test instruments in addition to the performance test instruments used to test the motor performance.

---

**Test and measurement of psychomotor skills**

Psychomotor skills are the combined intellectual and motor skills required to perform a procedural task.

Any motor performance has associated cognitive information processing components (intellectual skills). Most equipment operator and maintainer behaviors are psychomotor skills that are composed of one or more cognitive processes (intellectual skills) and one or more procedural steps and step actions (motor skills).

---

## Bibliography

- Bills, C.G., and Butterbrodt, V.L. (1992). *Total Training Systems Design Function: A Total Quality Management Application*. Wright-Patterson AFB, Ohio.
- Briggs, L.J., and Wager, W.W. (1981). *Handbook of Procedures for Design of Instruction* (2nd Ed.). Glenview, Illinois: Harper Collins Publishers
- Carlisle, K.E. (1986). *Analyzing Jobs and Tasks*. Englewood Cliffs, New Jersey: Educational Technology Publications.
- Davies, I.K. (1976). *Objectives in Curriculum Design*. London: McGraw Hill.
- Dick, W., and Carey, L. (1990). *The Systematic Design of Instruction* (3rd Ed.). Glenview, Illinois: Harper Collins Publishers.
- Gagné, R.M. (1985). *The Conditions of Learning* (4th Ed.). New York: Holt, Rinehart and Winston.
- Gagné, R.M., Briggs, L.J., and Wager, W.W. (1992). *Principles of Instruction* (4th Ed.). New York: Harcourt Brace Jovanovitch College Publishers.
- Gagné, R.M., and Merrill, M.D. (1990). *Integrative Goals for Instructional Design*. Englewood Cliffs, New Jersey: Educational Technology Publications. 38(1), 1-8.
- Goldstein, I.L. (1986). *Training In Organizations: Needs Assessment, Development, and Evaluation* (2nd Ed.). Pacific Grove, California. Brooks/Cole Publishing Company.
- Hageman, D.C. (1988). *Cognitive Engineering of Training Systems for Simulators*. National Aerospace and Electronics Conference. Dayton, Ohio.
- Hageman, D.C. (1985). *Effective Training Systems for High-Technology Equipment Operation*. National Security Industrial Association Fifth Annual Conference on Personnel and Training System Effectiveness. San Antonio, Texas.
- Keller, J.M. (1987). The Systematic Process of Motivational Design. *Performance and Instruction*, 26(9), 1-8.
- Kibler, R.J. (1981). *Objectives for Instruction*. Boston: Allyn and Bacon.
- Knirk, F.G., and Gustafson, K.L. (1986). *Instructional Technology: A Systematic Approach to Education*. New York: Holt, Rinehart, and Winston.
- Leshin, C.B., Pollock, J., and Riegeluth, C.M. (1992). *Instructional Design Strategies and Tactics*. Englewood Cliffs, New Jersey: Educational Technology Publications.
- Mager, R.F. (1962). *Preparing Objectives for Instruction* (2nd Ed.). Belmont, California: Fearon Publishers.
- Merrill, M.D., Tennyson, R.D., and Posey, L. (1992). *Instructional Design Strategies and Tactics*. Englewood Cliffs, New Jersey: Educational Technology Publications.
- Merrill, M.D., Lee, Z., and Jones, M.K. (1990). *Second Generation Instructional Design*. Englewood Cliffs, New Jersey: Educational Technology Publications.
- O'Neil, H.F., Jr., and Baker, E.L. (1991). Issues in Intelligent Computer-Assisted Instruction: Evaluation and Measurement. In T. Gutkin and S. Wise (Eds.), *The Computer and the Decision Making Process*. Hillsdale, New Jersey: Erlbaum Lawrence Associates.
- Reigeluth, C.M. (1983). Instructional Design; What is it and Why is it? In C.M. Reigeluth (Ed.), *Instructional Design Theories and Models? An Overview of Their Current Status*. Hillsdale, New Jersey: Erlbaum Associates.
- Rossett, A. (1987). *Training Needs Assessment*. Englewood Cliffs, New Jersey: Educational Technology Publications.

**Bibliography (Continued)**

Spears, W.D. (1983). *Processes of Skill Performance: A Foundation for the Design and Use of Training Equipment*. (NAVTRA EQ-VIPCEN 78-C-0113-4). Orlando, Florida: Naval Training Equipment Center.

Tennyson, R.D., and Michaels, M. (1991). *Foundations of Educational Technology; Past, Present and*

*Future*. Englewood Cliffs, New Jersey: Educational Technology Publications.

Wolfe, Petal. (1991). *Job Task Analysis: Guide to Good Practice*. Englewood Cliffs, New Jersey: Educational Technology Publications.

## Chapter 2

# GUIDELINES FOR CRITERION-REFERENCED TEST AND MEASUREMENT DEVELOPMENT AND USE

---

### Purpose of this chapter

The information in this chapter is to be used in conjunction with the information contained in AFM 36-2234, Instructional System Development, and in AFH 36-2235, Information for Designers of Instructional Systems, Volumes 1-11.

The purposes of this chapter are to:

Describe the concepts of test validity and reliability, retention and transfer tests, common types of predictive written test items, the validity of predictive written tests, performance test and measurement, and process rating methods.  
Provide guidelines for development of retention and transfer tests, predictive written tests, and performance tests and rating scales.

---

### Where to read about it

This chapter contains nine sections.

Section	Title	Page
A	Introduction	29
B	Test Validity	31
C	Test Reliability	33
D	Retention and Transfer Tests	37
E	Overview of Predictive Written Tests	46
F	Common Types of Predictive Written Test Items	48
G	Validity of Predictive Written Tests	51
H	Performance Tests and Measurement	54
I	Types of Process Rating Methods and Ways to Avoid Rating Errors	60

---

## Section A Introduction

---

### Purposes of tests

Tests can serve four purposes:

Measures for passing or failing students.  
Indications of whether a class is up to standard in specific cognitive and psychomotor skills.  
Indications of instructor proficiency.  
A means of diagnosing and correcting problems or weaknesses in the instructional system.

---

### Test validity and reliability

Tests must be valid and reliable.

*A valid test* measures what it's supposed to measure.  
*A reliable test* yields consistent results.

---

### Kinds of test feedback

Tests provide two kinds of feedback:

Whether the student retains specifically what is learned during instruction.  
Whether the student can transfer what is learned during instruction to the job.

---

### Forms of tests

Tests are either Predictive Measures or Performance Measures:

Predictive tests can be reliable and valid measures of enabling objective intellectual or motor skills.  
Performance tests can be reliable and valid measures of terminal or enabling objective intellectual or motor skills.  
Predictive tests are most useful for developing a direct measure of intellectual or motor skills specified in enabling objectives.  
Performance tests are especially suited to evaluating either a process (such as a procedure) or a product (such as an assembled piece of equipment) and the intellectual skills specified in terminal or enabling objectives.

---

---

**Overview of the  
test development  
process**

Follow these guidelines for test development:

Review your resources to ensure that you can develop tests for all the objectives in the course of instruction.

Determine the best type of test item for each objective.

Develop test items for all the intellectual and motor skills in each objective.

Develop *retention test items* for intellectual and motor skills that have been covered in the instruction.

Develop *transfer test items* for intellectual and motor skills that have not been covered in instruction.

---

## Section B

### Test Validity

---

**Definition of test validity**

Test validity is the relevance of a test to its purpose. *Validity is the most important characteristic of a test.* For Criterion-Referenced tests, validity refers to two characteristics of test items:

The extent to which test items are direct reflections of the objectives.

The adequacy with which the test items sample the objectives.

---

**Test items as reflections of objectives**

Preparation of Criterion-Referenced objectives simplifies construction of Criterion-Referenced tests. Criterion-referenced objectives describe:

The conditions of intellectual or motor skill performance.

The intellectual or motor skill performance required of the student after instruction.

The accuracy and/or time standards for intellectual or motor skill performance.

---

**Validity requirements of test items**

A test item is valid when:

It requires the student to demonstrate the intellectual or motor skill performance stated in the objective.

It requires that intellectual or motor skill performance be performed under the conditions stated in the objective.

It is scored according to the intellectual or motor skill standards stated in the objective.

---

**Using adequate sampling**

Not only must each test item be valid, but the test itself must be valid:

The validity of the entire test depends on how well its items sample the objectives.

A test which samples one small component of a course or unit of instruction to the exclusion of the rest of the instructional components is not a valid test.

---

---

**Assuring valid sampling**

To assure a valid sampling of test items and to keep the length of the test reasonable:

Separate the objectives into two groups, the critically important and the less important.

Include more than one test item for each critically important objective, and at least one test item for the less important objectives.

Create a pool of test items and rotate the test items used in a course of instruction.

---

**Validity of the instructional system**

There are two types of validity for an instructional system:

Instructional Validity: The degree to which the students learn during exposure to the instructional system.

Transfer Validity: The degree to which what has been learned in the instructional system transfers as enhanced performance to the job environment.

---

## Section C Test Reliability

---

**Definition of test reliability**

“Reliability” is the consistency with which a test measures student mastery of the instructional objectives.

If a criterion test is reliable, students who have mastered the objectives will always pass, and those who have not will always fail.

If a criterion test is unreliable, a student may pass or fail for reasons other than the ability to master the objectives.

---

**Methods for determining test item reliability**

Various methods have been devised for obtaining numerical indices of test and test item reliability.

Most are not suitable for Criterion-Referenced tests.

Those that are suitable are seldom practical.

The methods involve giving the same test to the same students on two different occasions, and comparing scores to see if the students did about the same in both instances.

It is often impossible to test the same people twice.

When the test is administered the second time, the scores may be affected by the varying amounts that different students learned when taking the first test or during the interim period of time between the first and second test administrations.

---

**Main factors in criterion-referenced reliability**

The four main factors in Criterion-Referenced test reliability are:

The test itself, including general and specific test instructions, and the conditions under which the test is administered.

The student taking the test.

The scoring procedures.

The length of the test.

---

---

**Reliability in  
criterion-  
referenced test  
administration**

The following guidelines increase reliability for Criterion-Referenced tests:

Give the test under the most consistent conditions possible. This is the most general principle of test administration. To illustrate, suppose the national champion high school runner was chosen by having students all over the United States run once around their own outdoor high school track on 1 December. Would this test be given under consistent conditions? Tracks would be different lengths. Track surfaces would vary from grass to concrete. Tracks could be different sizes. Weather conditions could include rain, snow, sleet, or dry. The winner of the competition may not be the fastest runner, but rather the one who ran under the best conditions. Make instructions to the student as clear and simple as possible. Criterion-referenced tests are not supposed to be tests of the student's ability to understand complex directions. Tell the student how the test will be scored. Inform the student whether speed or accuracy is more important. Inform the student if there are penalties for errors, or if the test will give the student simple credit for correct answers. Write all instructions, and make directions as complete as possible without giving away answers to test items. Decide in advance how much information is to be given to the student, and include this information in written instructions.

---

**Reliability in  
criterion-  
referenced test  
administration for  
test administrators**

Provide the test administrator with complete written instructions on all phases of test administration. These instructions should cover:

How to treat the students.  
What student questions can be answered.  
What equipment and supplies are needed for the test, and how these should be laid out.  
What to do in the event of various circumstances, such as student illness, equipment failure, or severe weather.

---

---

**Other factors important to reliability of test administration**

Other factors to consider:

Provide for thorough training of the test administrator. The administrator must provide adequate supervision to ensure that tests are given as prescribed.

Make sure that adequate supplies are available, and that equipment is in good working order. Inconsistency in test results will occur if these factors are not addressed.

Inspect and calibrate the equipment and tools used for testing frequently to ensure consistency of operation.

Protect students from extremes of environmental conditions which might affect test scores.

---

**Factors related to student reliability**

The student may be a source of unreliability. Illness, fatigue, the stress of the test, and lack of motivation may contribute to poor test scores even if the student has mastered the objectives.

The student should be rested, and treatment during the test should be designed to prevent the student from becoming excessively afraid of failure.

---

**Consistency of test scoring**

Scoring of tests is a major source of inconsistency. Scoring must be consistent from student to student.

---

**Objectivity**

The key principle to observe in scoring is objectivity. Objectivity is achieved by:

Setting precise standards, and training the test administrator to apply them.

Developing scoring procedures in which subjective judgment or opinion of the scorer is not a factor.

Telling the test administrator exactly what should be observed while scoring.

Clearly stating the standards of performance.

Defining successful performance so that measurements do not depend on personal judgments.

---

---

**Specifying standards**

Specifying standards is essential to objectivity and reliability. The following should be specified:

Specifying standards for intellectual skills based on a single correct answer.

Specifying standards for psychomotor training requirements.

Specifying standards that indicate if a student “did” or “did not” do a particular thing.

Specifying standards that indicate if a product exhibits the presence or absence of essential attributes.

Specifying standards that indicate if a procedure is performed within specific numerical parameters.

---

**Other ways to improve reliability**

Ensure that measuring instruments are accurate and calibrated.

Validate scoring procedures by having several scorers score one student.

Identify the reason for any differences in scores.

Make the standards more specific to correct differences in scores.

---

**Length of tests and test reliability**

If all other factors that influence test reliability are under control, you can improve the reliability of a test by making it longer.

If several test items for the same objective are included, the effects of wording of the instructions and scoring tend to cancel, and the overall score becomes more reliable.

If testing time is limited, and there is a choice between adding more items to cover the same objectives, or covering new objectives with new items, it is better to cover new objectives with new items.

---

## Section D

### Retention and Transfer Tests

---

#### Overview

It is possible that a student might pass a test and still not accomplish the education or training requirement. This could happen if either the instructional program or the test was inadequate.

---

#### Differences between retention and transfer tests

The test could be valid, in that it measured how well the student *retained* the specific course material, but not how well the material is transferred.

For example, a student who remembered how to solve a particular problem in class would pass a test item requiring solution of the same problem.

The test would measure *retention* of course content, but the student might not be able to solve new problems on the job.

The test has not measured how well the student *transfers* what has been learned to the job.

The following table describes the important differences between retention and transfer tests.

Retention Test	Transfer Test
Requires the student to demonstrate the retention of knowledge and skills acquired during instruction. The <i>same</i> examples and situations experienced in instruction are included in the test. The student must remember what was encountered during instruction.	Requires the student to demonstrate the retention of knowledge and skills acquired during instruction and the ability to apply them to <i>new</i> situations and examples not encountered during instruction. Different (novel) examples and situations are included in the test.

---

### Behaviors and testing

The following table indicates the relationship between the types of behavior and possible testing for retention or transfer.

Type of Intellectual Behavior (From education or training objective)	Type of Test	
	Retention	Transfer
Associating (Facts)	X	
Chaining (Motor or Verbal)	X	X
Discriminating	X	
Classifying	X	X
Rule Using	X	X
Problem Solving	X	X
Verbal Information	X	X
Cognitive Strategies/Metacognition	X	X
Motor Skills	X	X

### Testing for retention

Strict security should be maintained for *transfer* tests to prevent students from practicing in advance, or to prevent instructors from teaching the test.

For *retention* tests, teaching the test is not a problem. There may be only one correct way to perform the task. In these cases, it is fine to “teach the test”. Give the students the objectives at the beginning of the course.

Retention tests require the student to remember something presented in the instruction. Tests requiring remembering can take three forms as illustrated in the following table:

**Testing for retention  
(Continued)**

Memorization	Recall	Recognition
<p>A test item requires the student to write, state or perform in exact terms. The student must memorize exactly the content of the instruction. Any deviation is considered an error.</p>	<p>A test item may require the student to paraphrase or approximate what has been taught during instruction.</p>	<p>A test item may require the student to look at or read alternatives and recognize the correct answer. The correct answer has been encountered during instruction.</p>
<p>Test Item Examples: Write the formula for water. State the steps for removing the fuel pump.</p>	<p>Test Item Examples: In your own words, define the term discrimination. Demonstrate an acceptable method for starting a car.</p>	<p>Test Item Examples: Which of these two fuel pumps are correctly assembled? Select the correct formula from this list.</p>

**Testing for transfer**

Transfer tests require the student to memorize, recognize or recall several intellectual or motor skills mastered during instruction and to apply these skills to new (novel) situations not encountered during instruction.

For example, the student may have to *use learned rules to solve novel problems* requiring the use of a formula or using specific procedural steps.

You can't test for transfer if the student has access to the test items and "learns" only those problems on the test.

The student should have practiced on typical problems of this sort prior to administration of the transfer test.

The whole purpose of a transfer test is to see if the student can apply learned intellectual or motor skills to novel conditions.

**Sampling of complex behaviors**

Transfer tests can be used to measure complex psychomotor skills.

For example, in teaching a pilot to land a plane, it is not feasible to use all possible landing strip configurations. A good transfer test would sample from the various classes of landing strip configurations to measure a student's ability to transfer learned psychomotor skills to conditions not encountered in training.

**Types of transfer test items**

The three primary types of transfer test items are:

Recognition  
Production  
Application

<b>Summary of the Three Types of Transfer Test Items</b>		
<b>Recognition</b>	<b>Production</b>	<b>Application</b>
A test item requires the student to look at or read alternatives never encountered in instruction, and to <i>recognize the correct answer.</i>	A test item presents the student with a novel practical example or situation. It asks the student to <i>state or produce the correct answer or procedure.</i>	A test item presents the student with a novel practical problem. It asks the student to <i>solve the problem using principles or procedures not encountered in instruction.</i>

**Types of transfer test items  
(Continued)**

Examples of Recognition Test Items	Examples of Production Test Items	Examples of Application Test Items
<p>Which of the following (new) examples represent negative reinforcement? Read the statement and select the specific answer that describes the statement.</p>	<p>Give an example of negative reinforcement not discussed in class. Read the case study and state the specific disorder which describes the patient. Select the best strategy for handling the mental patient described in the study. Troubleshoot an equipment malfunction not specifically covered in instruction.</p>	<p>Read this case study of a mental patient, and using principles of reinforcement, generate a resource utilization strategy for managing the patient. Generate tactics for landing an aircraft under conditions not encountered in instruction. Perceive job performance condition cues, and generate judgments as to whether a cue is an indicator of an abnormal or emergency condition, and the probable cause of the condition.</p>

**Retention or transfer test items?**

Whether you test for retention or transfer depends on the kind of behavior involved in the instructional objective.

Retention tests use *Memorization, Recall, or Recognition* test items.

Retention tests are used to measure mastery of intellectual or motor skills contained in the course of instruction.

---

**Retention or transfer test items?  
(Continued)**

Transfer tests use *Recognition, Production, or Application* test items.

Transfer tests require the student to memorize, recognize or recall several intellectual or motor skills mastered during instruction and to apply these skills to new (novel) situations not encountered during instruction.

---

**Overview of transfer test development**

To develop a transfer test for concepts mastered during instruction:

Develop a list of examples and non-examples of each concept taught in the course of instruction.

The number of these examples to use in the test is based on the difficulty the students have in learning the concept.

---

**Testing concepts**

Concepts have the following characteristics:

Concepts include a *class* of people, events, objects, or ideas. Members of a class share some *common properties or attributes*.

The individual members of a class are clearly different from each other on *some* properties or attributes.

Concepts have many examples of application. It is impossible to teach them all.

To test a concept, create examples that use the concept, and then select a sample of the examples to use in the test.

---

**Examples and non-examples of a concept**

An *example* has the essential attributes of the concept.

For example, for the concept “round”, rolling is an essential attribute. Since a ball rolls, it is an example of the concept “round”.

A *non-example* lacks the essential attributes of a concept, although it may share some irrelevant attribute with other members of the class.

Suppose all round objects presented to teach the concept “round” happened to be red. A red ball would be an example of “round”, not because it’s red, but because it rolls. A red cube would be a non-example of round — it’s red, but it doesn’t roll.

---

---

**Testing for transfer of a concept**

When testing for transfer of a concept:

Ensure that students correctly make the *same* response to a new member of the class, which differs in some way from previously used examples of the class members (For example, if one round object shown during instruction was a phonograph record, a test item might include another example, such as a dinner plate).

Ensure that students correctly make a *different response* to non-examples which share some incidental attributes with the members of a class (For example, if all the round objects presented in instruction were red, a test item might include a non-example of a red cube).

Use examples and non-examples during instruction, and in the Criterion-Referenced test.

---

**Advantages of using examples and non-examples**

Using examples and non-examples during instruction will help the student learn to avoid two common problems:

The student will learn to include all true examples as members of the class, and will be better able to transfer what has been learned to the job environment.

The student will learn to exclude non-examples from membership in the class, and will be better able to transfer what has been learned to the job environment.

---

**Selecting examples and non-examples**

Base your selection of examples and non-examples on the attributes of the members of the class of concepts, principles, etc. Some attributes are critical (round objects roll). Other attributes are incidental (round objects come in various colors).

To prepare a list of examples and non-examples of a concept:

Determine the critical attributes shared by all members of the class.

Determine the incidental attributes which might lead students to make errors. (These are properties of the members of a class that could cause a student to incorrectly classify a non-example as an example.)

---

---

**Selecting examples and non-examples (Continued)**

Prepare a list of examples and non-examples:  
Use enough examples to vary each incidental attribute.  
Use enough non-examples to exclude each critical attribute.

Select from the total list of examples and non-examples those that will be used in testing for transfer.

---

**Sampling from a list of examples and non-examples**

To select a sample of examples and non-examples from a prepared list of examples and non-examples of a concept:

*Determine the size of the sample:* Determine how large a sample is needed to test for transfer. The size of the sample depends on how difficult the concept is to learn.

---

**Factors in transfer test development**

Many factors contribute to the difficulty of learning a concept, however three are particularly relevant for developing an adequate transfer test:

The number of members of a class.

The number of critical attributes that could be used to describe each member of the class.

The similarity of the critical and incidental attributes.

---

**Number of members of a class**

Determine the number of members of a class:

If student performance requires distinguishing among a large number of members in a class, sample more heavily than for a class having only a few members.

The more members there are in a class, the harder it is to see the essential similarities between them. A large class could have a dozen members.

---

**Number of critical attributes of each member**

Determine the number of critical attributes of each member:

The larger the number of critical attributes the student must know, the harder it will be for the student to see the essential similarities among the members of the class.

For example, it is harder to classify objects on the basis of size, shape, color, and texture than on the basis of color alone. When there are more than three critical attributes, you should sample more heavily.

---

**Similarity of critical and incidental attributes**

Determine the similarity of critical and incidental attributes:

The more similar the critical and incidental attributes are, the more difficult it will be for students to identify only the correct members of the class.

When critical and incidental attributes are similar, you must sample both examples and non-examples heavily. If critical and incidental attributes are dissimilar, you can sample less heavily.

**Example of difficulty factors in learning a concept and the associated sample size**

The astronauts learned to classify minerals according to type. Suppose one objective required classifying minerals as quartz. To correctly classify sample minerals, the astronauts must understand the concept of "quartzness". The concept involves many different kinds of quartz (members of the class). There are several critical attributes as well. These include luster, hardness, streak, and specific gravity. The critical and incidental attributes are fairly dissimilar. (For example, the color of quartz, an incidental attribute, is not similar to any of the critical attributes).

The following table depicts the difficulty factors in learning a concept and the associated sample size.

<b>Difficulty Factors</b>			
<b>Number of Members in the Class</b>	<b>Number of Critical Attributes of Each Member</b>	<b>Similarity of Critical and Incidental Attributes</b>	<b>Number of Examples and Non-examples to Sample</b>
Few (<5)	Few	Dissimilar	Few
Few	Several (<5)	Dissimilar	Many
Few	Few	Similar	Moderate (5-10)
Few	Several	Similar	Many
Many (>10)	Few	Dissimilar	Few
Many	Several	Dissimilar	Moderate
Many	Few	Similar	Moderate
Many	Several	Similar	Many

## Section E

### Overview of Predictive Written Tests

---

#### **Testing lower-level intellectual skills**

The acquisition of lower-level intellectual skills (Associations, Discriminations, Concrete and Defined Concepts, and Rule Learning) can be tested by paper and pencil or computer-generated test items.

Such items are usually printed or displayed in a specific format, with self-contained directions.

The student records answers on the test itself or on a special answer sheet, or interacts with a computer display to record the answer.

By performing a statistical analysis, the validity of a test or test items for prediction of hands-on job performance can be determined.

Interactive courseware (ICW) predictive tests (and some forms of paper and pencil tests) may be supported by graphic, video, photographic, audio, or reference materials with which the student is required to interact during the test.

---

#### **Degree of objectivity for predictive written tests**

Objective predictive written test items permit reliable scoring.

Objective test items can be accurately scored as correct or incorrect by a computer or an individual.

Objective predictive written test items have the advantage of being reliable instruments.

Objective predictive written test items are usually a less valid measure of job performance than a hands-on performance test.

Objective predictive written tests should be administered prior to evaluation of hands-on performance to ensure that all of the lower-level intellectual skills associated with a hands-on task are measured.

Subjective test items (such as ratings) make scoring unreliable, because scoring depends upon the judgment of experts who may have varying opinions about correctness.

Predictive written test items are used to predict the ability of a student to perform an actual job by measuring the intellectual skills that are related to or associated with actual job performance.

---

**Objective versus subjective test items**

To some extent, the degree of objectivity of a predictive written test item depends upon the type of test item as shown in the following table.

<b>Objective</b>	<b>Less Objective</b>	<b>Subjective</b>
Multiple Choice Matching Completion, when answers are short, requiring a specific phrase. True-False when indisputably factual or not factual.	Production items (case studies or problems) when answers are specific and not open to interpretation.	Essay in which a student is required to discuss a topic. Completion answers when answers can be phrased in various ways. True-False when dependent on context.

## Section F

### Common Types of Predictive Written Test Items

---

**Multiple choice test items** The multiple-choice question is appropriate for measuring most lower-level intellectual skills such as discrimination, concrete concepts, defined concepts, and rule learning. It is versatile, and can be used to measure facts, terminology, concepts, and principles and can take various forms.

The test item contains the best answer.

The test item can measure retention or transfer.

The best answer does not have to be the one and only indisputably correct answer. However it must be defensible as the most nearly correct to enable measurement of student comprehension of the alternatives presented.

---

**Matching test items** Matching items typically use two columns of related words, phrases, symbols, or illustrations. The student matches each element in one list with the most closely related element in the other list.

If the number of alternatives in one list is the same as the number of items in the other list, the trainee who knows the answers to all but one of the items will automatically get this one correct also. This problem can be avoided by including two or three more items in the list of alternatives than there are items. It can also be avoided by having the same alternative be the answer for more than one item.

Avoid the use of alternatives that can be correctly paired or rejected without any knowledge of the subject, as in the following example:

Match the type of Technical Order to the descriptive statement:

- |  |                  |
|--|------------------|
| 1. A listing of technical orders by number and title.    | A. TCTO          |
| 2. Identified by the word PRELIMINARY on the title page. | B. Preliminary   |
|  | C. Built-In-Test |
|  | D. Index         |
-

---

**Matching test items  
(Continued)**

3. Inspects and tests equipment automatically.
4. Directs when work will be accomplished.

Matching items can measure retention and transfer of lower-level intellectual skills such as discrimination, concrete concepts, defined concepts, and rule learning, and take many forms.

---

**Completion test items**

Completion items may take two forms:

A question may be asked that requires only a single word or phrase to answer.

A sentence may have one or more internal blanks to be filled in.

In either case, scoring is more objective when the answer is a specific word or phrase.

Completion test items can be used to measure retention and transfer of lower-level intellectual skills such as discrimination, concrete concepts, defined concepts, and rule learning.

Oral completion tests may be used to measure student retention and transfer of higher-level intellectual skills such as declarative knowledge (verbal information), the stating of cognitive strategies or tactics in response to operational conditions, monitoring the use of strategies and tactics (metacognition), or attitudes and motivation.

One abuse of completion test items is to carry them to extremes with excessive blanks. For example: "\_\_\_\_\_ test items may be \_\_\_\_\_ and \_\_\_\_\_ because \_\_\_\_\_."

---

**True-false test items**

True-false test items should not be included in Criterion-Referenced tests. True-false test items have several disadvantages:

Verbatim statements are frequently lifted from the instructional materials, with some negative terms

---

---

**True-false test items  
(Continued)**

included to make some items false. True-false test items encourage rote memorization by the student. Because statements are out of context, it may be difficult to defend many items as undeniably true or false. Therefore, most true-false test items are factual in nature. This again leads to the criticism that true-false items over-emphasize the memorization of facts. Because true-false test items have only two possible answers, students can be expected to get half of them correct by guessing. Even if you statistically correct for guessing, it is difficult to establish a meaningful criterion performance level.

---

**Production test items**

Avoid using written production test items in Criterion-Referenced tests that measure lower-level intellectual skills Associations, Discriminations Concrete/Defined Concepts, and Rule Using.

Oral production test items are used to measure acquisition of higher-order intellectual skills (Verbal Information, Cognitive Strategies, and Metacognition)

It is difficult to score production test items objectively, unless specific metrics are established for each test item.

Written production items may be useful for testing for transfer of intellectual skills to a condition not encountered in training. However, equally effective multiple-choice and matching test items can be written that test and measure the same transfer of intellectual skills.

---

**Increase objectivity of production test items**

Increase the objectivity of production test items by:

Preparing a key to minimize subjectivity as much as possible. Include all important details such as key words, phrases, and response time metrics for each test item.

Having the evaluators use code numbers for student responses to decrease rater bias.

---

## Section G

### Validity of Predictive Written Tests

---

#### Advantages of predictive written tests

Objective predictive written tests offer numerous advantages:

They can be reliably administered.

They can be machine-scored.

They can cover a large amount of material in a short period of time.

Test score data are easily maintained for record-keeping purposes.

Statistical data describing certain test item characteristics such as difficulty, the mean and variance of test items, correlation between test items, response patterns, internal test consistency, and test variance can be computed.

---

#### Validity of types of predictive written test items

The validity of predictive written tests is a major problem. Even if the learning objectives relate directly to job performance, it is often difficult to relate predictive written tests directly to job performance. Consider these examples for different types of test items:

Multiple-Choice test items require the student to discriminate between several possible answers to select the correct one (recognition test item).

---

#### Multiple choice test items

The only type of job objective for which the multiple-choice test item is directly suited is one that requires the student to select from among several alternatives on the job.

Examples of selection from a set of alternatives on the job could be:

Selection of tools for a given purpose.

Selection of a proper procedure from several described in a Technical Manual.

The multiple-choice test item is inappropriate for most job situations where the student must do something (psychomotor behavior), such as disassemble a weapon, perform a flight maneuver or procedure, or make a maintenance repair.

---

---

**Matching test items** Matching items might appear to be appropriate for testing associations between concepts, but there can be problems with validity of the test items.

---

**Completion test items** For example, a student could match a list of telegraphic code dots and dashes with the appropriate letters. This is not a directly valid measure, since the job requires translating an audible code, not a visible one, into a written message.

Completion test items are useful for testing an item of knowledge in a specified context.

Completion test items are appropriate for testing intellectual skills such as discrimination, concrete concepts, defined concepts, and rule learning.

Completion test items are appropriate for testing associations, some discriminations, portions of chains (fill in the missing steps) and intellectual skills related to more complex types of behavior, such as declarative knowledge (verbal information).

Oral completion tests may be used to measure student retention and transfer of higher intellectual skills such as declarative knowledge (verbal information), the stating of cognitive strategies or tactics in response to operational conditions, monitoring the use of strategies and tactics (metacognition), or attitudes and motivation.

There may be validity problems even with completion test items. The direct relevance to job performance may be questionable.

For example, a completion item could require a student to list the four major distinguishing features of quartz. The student may pass the item from memory, but does that mean that the student can always identify quartz when given a variety of mineral samples?

---

**True-false test items** True-false items can be used for testing the memorization of a specific fact, or for testing simple discriminations.

True-false test items have several disadvantages that were mentioned in Section F.

The relevance of true-false questions to job performance is questionable.

---

---

**Production test items**

Production test items can be useful for listing the steps of a procedure (chaining), or describing a particular discrimination, classification, rule, (or approach) to solving a problem.

Because of their limitations (Section F), ensure that production test items are good measures of job performance.

Predictive written test items are especially useful for assessing the intellectual skill components of a task.

Ensure that the test items are valid.

---

**Direct validity of test and job performance**

Intellectual skill tests are indirectly related to job performance.

Test items must be statistically compared with a *directly valid measure of job performance*.

If a test item is statistically correlated with a valid measure of job performance, the test item can be described as an *indirectly valid* measure of job performance.

The indirect validity of test items can be determined by the degree of correlation between student performance on a test item and actual job performance related to the test item.

Predictive written test items should be statistically correlated with job performance measures to ensure indirect validity of each test item.

---

**Indirect validity**

One method to determine the indirect validity of a test is to administer the test to a group of experts and to a group of naive students.

If the experts do well on the test and the students do not, the test has indirect validity.

If the experts and the students both do well on the test, there is no need for instruction (if the test is a valid measure of criterion performance).

---

## Section H

### Performance Test and Measurement

---

#### General characteristics of performance tests

*Performance tests* require the student to do something (*perform a psychomotor skill behavior*). *Predictive tests* require the student to verbalize, answer questions, or write about something (*perform an intellectual skill behavior*).

---

#### Comparison of performance and predictive test items

The following table summarizes the major differences between performance and predictive tests.

Performance Test Items	Predictive Test Items
<p>Requires the individual to accomplish a job-like task under controlled conditions. Emphasize non-verbal aspects.</p> <p>May require individuals to look up, read, and use certain job aids.</p> <p>Test items are the psychomotor skills the individual must perform, and the judgments and decisions that must be made to perform the procedure under normal and abnormal job performance conditions.</p> <p>Test items are dependent upon the sequence in which they are presented. Errors made early in the sequence may affect the final outcome of the task.</p>	<p>Requires the individual to demonstrate intellectual skills by responding to various types of written or oral questions.</p> <p>Emphasize verbal or symbolic aspects.</p> <p>May require individuals to look up, read, and use certain reference materials.</p> <p>Test items measure job-related intellectual skills that the individual must know to perform a job or to make judgments or decisions during job performance.</p> <p>Test items are usually independent questions, and are not dependent upon sequence. Errors on one item will not usually affect performance on another item. (Some test items can be dependent, and errors made early in the sequence can affect the final score on the test.)</p>

---

---

**Direct and indirect performance measurement**

Often it is difficult in a training situation to completely reflect the job situation in performance tests.

The best performance test environment is one which closely approximates a typical real-life situation.

Practical situations may make it necessary to settle for something less than the best performance test environment. For example, an attempt to measure an individual's ability to drive an automobile would require a performance test that would measure the student's ability to perform all driving tasks of all automobiles, on all types of roads, in all traffic conditions, and under all types of weather conditions. Obviously, it would be impossible to meet all of these conditions under a practical performance test environment.

If a predictive test is used to measure the performances that cannot be measured in the test environment, they cannot be considered to be a valid substitute for measurement of performance unless a high empirical relationship to the Criterion-Referenced psychomotor performance can be demonstrated.

---

**Validity of performance tests**

A student's ability to perform a task is evaluated by observing student behavior and judging it against a pre-defined Criterion-Referenced standard.

Performance test items rate the student on the actual performance output defined in a specific performance objective.

If the specific output is a motor skill (or process), the student is required to display the *performance*.

If the specific output is a product, the student is required to produce the *product*.

Performance tests directly measure training requirements, and are inherently *more predictive of job performance*.

Performance tests usually require judgment by an test administrator.

---



---

**Rating scales and their use in criterion-referenced performance measurement (Continued)**

The rating scale can be used to provide feedback to the student on progress made towards achieving the Criterion-Referenced performance objective.  
A rating scale can help flag a need for revising the course materials. If many students get low ratings, there is a good possibility that the fault is with the course materials, and not the students.

---

**Common errors in rating performance**

Rating scales that measure criterion performance require test administrators to observe and score student performance. Since the scoring is based on judgments, observer ratings may be less reliable.

---

**Requirements of performance tests**

Performance tests require the student to display actual outputs (product or process).  
Performance tests depend heavily on actual observations and rating of outputs.

---

**Conditions for performance testing**

Each student should be tested under conditions that provide the best chance to display the skill or product.

The test conditions should not change from one student to another.

The test administrator must observe student performance and rate the performance according to a fixed standard.

1      2      3      4      **5**      6      7



“5” is the rating needed to pass criterion test

When scales are used to judge quality, observers may differ in their judgments. These differences are called rating errors. Rating errors can be classified into three broad groups:

- Error of Standards (Affects all individuals rated by an observer)
  - Error of Halo (Affects only certain individuals within a group)
  - Logical Error (Appears only when two or more traits of individuals are being rated)
-

---

**Error of standards** Some observers tend to overrate or underrate because of the difference in their personal standards.

Standards of physical measurement are fixed units such as inches, centimeters, ounces, or grams.

In rating with only mental standards for performance, there may be as many different standards used to rate an individual's performance as there are observers.

---

**Error of halo** Observers sometimes allow their rating of performance to be influenced by their general impression of the individual.

**What it is**

A performance rating may be formed on the basis of observations or knowledge extraneous to the performance being rated.

Using extraneous observations or knowledge would result in a shift of the rating.

This shift is called Error of Halo. Halo errors can be either favorable or unfavorable, and affect only certain persons that are rated. If the observer is favorably impressed, the shift is towards the high end of the scale.

**Types of Halo Errors**

An Error of Leniency is a type of Halo Error. It occurs when a rater favorably shifts a rating of a friend or a close acquaintance.

An Error of Stereotype is another type of Halo Error. It occurs when a rater shifts a rating of a person about whom they have some preconceived concept such as concepts involving a racial or religious group. These preconceptions influence observers.

Error of Halo results from the likes, dislikes, opinions, prejudices, and moods of raters.

**Detection of Halo Errors**

Error of Halo can be positively identified only when many competent and experienced observers rate a number of persons under identical conditions and the ratings of one of the observers consistently disagrees as indicated by a favorable or unfavorable shift in a rating.

---

---

**Error of halo  
(Continued)**

An Error of Halo may frequently go undetected, although it may be suspected. Usually only extreme cases are detected, even under controlled conditions.

Even when an Error of Halo has been identified, its reappearance cannot usually be predicted. It is the most difficult error to overcome.

---

**Error of logic**

A logical error may occur when two or more traits are being rated. It is present if an observer tends to give similar ratings to traits which don't necessarily go together.

For example, some observers may think that an industrious person is also efficient. Industrious persons may often be efficient, but not necessarily so.

The term "logical error" means that the traits are related in the mind of the observer.

The relationship may not appear to be logical to someone else.

Usually a person who exhibits this error is not aware of it.

---

## Section I

### Types of Process Rating Methods and Ways to Avoid Rating Errors

---

#### Types of rating scales

There are scales for rating a performance that is observable but transient. Some of the types of scales are the numerical scale, the descriptive scale, the graphic scale, or the checklist.

If at all possible, use the checklist. The checklist is derived directly from job performance requirements, and is the most reliable scale.

---

#### Checklist

A checklist is useful for rating ability to perform a specific set procedure. It is also a simple method of rating performance skills when your purpose is to see if students have reached a certain minimum level of performance.

The following example is a portion of a checklist rating form for instrument flying proficiency that is used by an observer to indicate whether the completion of each step was satisfactory or unsatisfactory.

Breaking a specific set procedure into many observable elements greatly reduces the Error of Standards.

#### **CHECKLIST**

INSTRUCTIONS: If the performance is satisfactory, place a + sign in the space provided. If the performance is unsatisfactory, place a - sign in the space.

1. Maintains constant heading within 5 degrees of course.
  2. Maintains constant altitude within 50 feet.
  3. Can make a timed turn within 10 degrees of a new heading.
  4. Can make a steep turn within 50 feet of altitude.
-

---

**Checklist  
(Continued)**

Reliability is usually high in checklist rating because of the nature of the decisions required.

A reduced number of choices available to the observer requires a reduced number of judgments that must be made by the observer.

The chance for either error or bias is greatly reduced when the choices are reduced to two (satisfactory or unsatisfactory).

Because of broad differentiations in the rating scores (pass/fail, satisfactory/unsatisfactory), the checklist is a comparatively reliable rating method.

---

**Numerical scale**

A numerical scale divides the specific performance set into a fixed number of points. The number of points on the scale depends on:

The number of differentiations required.

The ability of observers to differentiate.

For example, a squadron operations officer must find out which pilots are below criterion performance levels on a specific procedural set so on-the-job training can be provided to individuals who are performing below criterion levels.

A rating scale could be used to document performance levels, but a problem is to determine the number of points that the numerical scale should have.

---

**Number of scale  
points**

The number of points needed on a numerical rating scale will depend, in part, upon how well observers can differentiate.

Most people are able to make at least five differentiations.

Few trained observers can reliably make more than nine differentiations. As a result, most rating scales contain five to nine points.

---

**Narrative point scale example**

The following example shows a simple numerical scale for rating pilot ability.

INSTRUCTIONS: Place a check mark in the scale above the number that most accurately describes the pilot being rated.				
PROCEDURE: MAINTAINS CONSTANT HEADING WITHIN 5 DEGREES OF COURSE				
			<b>Criterion-Referenced Standard</b>	
1	2	3	4	5

**Descriptive point scale example**

The descriptive scale uses phrases to indicate levels of ability. The following example shows a simple scale for rating pilot ability. Five levels of ability are described.

INSTRUCTIONS: Place a check mark in the scale above the word that most accurately describes the pilot being rated.				
PROCEDURE: Maintains constant heading within 5 degrees of course				
			<b>Criterion-Referenced Standard</b>	
UNABLE	FAIR	GOOD	<b>EXCELLENT</b>	SUPERIOR

**Descriptive scale advantages**

The descriptive scale is more versatile than the numerical scale, because the degrees of excellence can be varied to suit the occasion.

For example if the squadron operations officer feels that all pilots satisfy criterion performance, but wants to know to what degree each pilot is better than satisfactory.

A numerical scale might be useful, except for the common feeling that the lowest numbers on the scale indicate inferior performance.

By using a descriptive scale, the operations officer gives his observers a frame of reference. The Criterion-Referenced standard is labeled SATISFACTORY.

**Descriptive scale advantages (Continued)**

INSTRUCTIONS: Place a check mark in the scale above the word that most accurately describes the pilot being rated. PROCEDURE: MAINTAINS CONSTANT HEADING WITHIN 5 DEGREES OF COURSE				
<b>Criterion-Referenced Standard</b>				
<b>SATISFACTORY</b>	GOOD	EXCELLENT	OUTSTANDING	SUPERIOR

The major disadvantage in using descriptive scales is a semantic one. An “excellent pilot” does not mean the same thing to all observers. Another disadvantage is that it is hard to select phrases that describe degrees of performance that are equally spaced. For example, in the descriptive scale above, some people could feel that there is less distance between “excellent” and “superior” than there is between “satisfactory” and “excellent”.

**Graphic scale**

The graphic scale is a combination of the numeric and descriptive scales.

In addition to a numerical scale, various adjectives are set below a continuous horizontal line. The line represents the range of the ability or trait being measured. In using the graphic scale, the user must consider not only the numerical range of the scale, but also the phrases that describe the various positions on the scale.

**Scale for judging a trait**

Example A, the observer is given instructions for judging the trait of “industry.” The observer is told to mark the scale after considering energy and application to duties, day in and day out.

These instructions help reduce errors of halo and improve objectivity and reliability. They also help the observer to consider and rate the same things about each person. The descriptive phrases below the scale, however, allow errors of standards to affect the rating.

**Scale for judging a trait (Continued)**

Phrases that describe observable behavior would help reduce the error of standards and the error of halo.

<b>Example A</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>
<b>Industry:</b> Consider energy and application to duties day in and day out.	Lazy	Indifferent	Diligent	Energetic	Untiring

**Scale for judging behavior**

Example B shows a graphic scale in which certain types of behavior are described for each point on the scale.

With most scales, the observer must not only observe, but must also evaluate the observation to form a rating. Generally, people can observe more accurately than they can evaluate what they have observed. The difficulty of evaluation increases errors of rating.

Whenever ratings can be based on observations alone, reliability is greatly improved.

The scale in example B requires the observer to record and evaluate the actions of the person being rated.

This type of graphic scale incorporates a great deal of objectivity.

If a trained rater observes accurately and records honestly when using this type of scale, all rating errors except error of halo should be eliminated.

The error of halo itself should be considerably reduced because of the objectivity built into the scale.

In constructing this type of scale, the developer must make sure that the behavior described for each point is actually an improvement over the point just below it.

In each case, distances between the points should appear to the observer to be about equal.

**Scale for judging behavior (Continued)**

<b>Example B</b>	1	2	3	4	5
<b>Cooperation:</b> Demonstration of willingness to work with others.	Creates friction	Indifferent to others	Gets along with most people	Harmonious team worker	Actively promotes harmony in working with others

**Alternate scale for judging behavior**

The scale in example C is similar to the scale in example B, except descriptive phrases are not provided for all points.

Many times observers feel that the rating should fall somewhere between two points.

Such a rating is facilitated by the use of this type of graphic scale.

The fuller descriptions of example C increase the likelihood that observed behavior can be pinpointed on the scale.

Generally, more detailed descriptions should contribute to better rating results.

<b>Example C</b>	1	2	3	4	5
<b>Initiative:</b> Action taken on own responsibility.	Slow to act, even when a decision is needed. Waits for others. Lets opportunities pass. Does not volunteer. Reticent.		Takes needed action without delay. Volunteers for some tasks. Undertakes all routine jobs without supervision. Dependable.		Anticipates needs. Works ahead and prepares for possibilities. Actively seeks opportunities. Eager.

**Example of a product rating method**

Since a product, unlike performance, is a very tangible thing, a product rating is more reliable than a process rating.

The following notional example shows a product scale for rating the ability of a pilot to fly a specific ground track in a flight simulator.

**Example of a product rating method (Continued)**

The Criterion-Referenced standard ground track is shown at A on the product rating form.

On completion of the simulator mission, the pattern that the pilot actually flew in the simulator is compared with the patterns on the scale.

From the comparison of the product to the scale, a rating is produced.

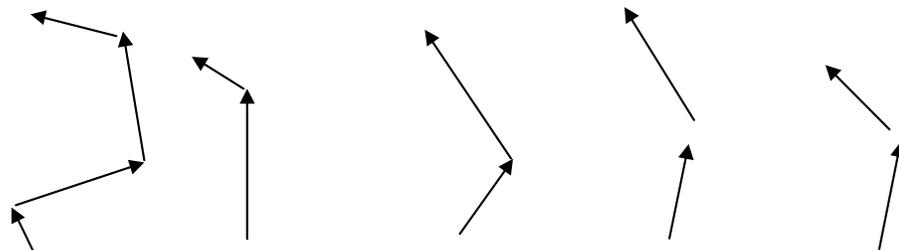
If followed carefully, this procedure can eliminate nearly all rating errors.

The scale provides a tangible standard that a rater can use to measure the product.

This type of scale eliminates errors of standard and errors of logic.

Halo error is not a problem, since the rater does not know who is responsible for the product.

SIMULATOR PRODUCT OF PERFORMANCE				
E	D	C	B	A
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<b>STANDARD</b>
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>



**Use of media for test and measurement**

Table 2 describes the use of types of media for test and measurement of intellectual skills and psychomotor performance.

Table 2 Use of Media for Test and Measurement

Type of Media	Description of Media Subsets	Test and Measurement Functions	Level of Test and Measurement Interactivity	Test and Measurement Fidelity
<b>Self Study</b>	Technical Manuals/Orders, workbook, audio/video tapes, to be used in conjunction with materials.	Cognitive: Predominately symbolic/ representational didactic process.	Representational. Abstract.	Not essential.
<b>Academics</b>	Classroom lecture, seminar, audiovisual media, mock-ups, demonstrations.	Cognitive: Concepts, procedural knowledge, decision-making knowledge.	Student/group responding activities, random/frequent verbal questions, frequent summaries.	Cognitive media generally restricted to the cognitive domain.
<b>Interactive Courseware (ICW)</b>	Computer-based instruction using interactive digital video, graphic, and audio media, or interactive videodisc media.	Cognitive and partial psychomotor. Perceptual skill development.	Situational simulation. Comprehension and application. Can be networked for team interaction.	Provides high functional fidelity and low physical fidelity.
<b>Familiarization Training Devices</b>	Equipment familiarization.	Cognitive: Knowledge of system operation. Procedural: Sequential operations.	Capability to interact realistically with the stimuli and response characteristics of procedural tasks.	Position of controls relative to their configuration and tactile characteristics. May not have total fidelity of real-time system operation.
<b>Part-Task Training Devices</b>	Equipment familiarization, normal, abnormal, and emergency procedures. Includes multi-task and unit training devices.	Cognitive: Knowledge of system operation. Procedural: Sequential operations. Psychomotor: tactile facilitation and stimulation. Decision-making knowledge.	Capability to interact realistically with the stimuli and response characteristics of specific procedural tasks.	Position of controls relative to their configuration and tactile characteristics. Total fidelity of real-time system operation may not be critical. Can be networked for team training.
<b>Operational Training Devices</b>	Simulator with or without visual system, domed simulator, networked systems.	Permits high skill development prior to, or in conjunction with, the actual equipment training phase.	Full scale.	Near-full to full-scale fidelity.

Table 2 Use of Media for Test and Measurement (Continued)

Type of Media	Description of Media Subsets	Test and Measurement Functions	Level of Test and Measurement Interactivity	Test and Measurement Fidelity
<b>Weapon System Training Devices</b>	Weapon system trainers for specific equipment.	Supports full mission training or rehearsal.	Full scale.	Limitations in field-of-view resolution and luminance.
<b>Actual Equipment</b>	Embedded training, electro-optical devices and helmet displays are options.	Supports full mission training or rehearsal.	Full scale.	Limited by safety considerations for personnel and equipment.

## Bibliography

- Bills, C.G., and Butterbrodt, V.L. (1992). *Total Training Systems Design Function: A Total Quality Management Application*. Wright-Patterson AFB, Ohio.
- Briggs, L.J., and Wager, W.W. (1981). *Handbook of Procedures for Design of Instruction* (2nd Ed.). Glenview, Illinois: Harper Collins Publishers
- Carlisle, K.E. (1986). *Analyzing Jobs and Tasks*. Englewood Cliffs, New Jersey: Educational Technology Publications.
- Davies, I.K. (1976). *Objectives in Curriculum Design*. London: Mc Graw Hill.
- Dick, W., and Carey, L. (1990). *The Systematic Design of Instruction* (3rd Ed.). Glenview, Illinois: Harper Collins Publishers.
- Gagné, R.M. (1985). *The Conditions of Learning* (4th Ed.). New York: Holt, Rinehart and Winston.
- Gagné, R.M., Briggs, L.J., and Wager, W.W. (1992). *Principles of Instruction* (4th Ed.). New York: Harcourt Brace Jovanovitch College Publishers.
- Gagné, R.M., and Merrill, M.D. (1990). *Integrative Goals for Instructional Design*. Englewood Cliffs, New Jersey: Educational Technology Publications. 38(1), 1-8.
- Goldstein, I.L. (1986). *Training In Organizations: Needs Assessment, Development, and Evaluation* (2nd Ed.). Pacific Grove, California. Brooks/Cole Publishing Company.
- Hageman, D.C. (1988). *Cognitive Engineering of Training Systems for Simulators*. National Aerospace and Electronics Conference. Dayton, Ohio.
- Hageman, D.C. (1985). *Effective Training Systems for High-Technology Equipment Operation*. National Security Industrial Association Fifth Annual Conference on Personnel and Training System Effectiveness. San Antonio, Texas.
- Keller, J.M. (1987). "The Systematic Process of Motivational Design." *Performance and Instruction*, 26(9), 1-8.
- Kibler, R.J. (1981). *Objectives for Instruction*. Boston: Allyn and Bacon.
- Knirk, F.G., and Gustafson, K.L. (1986). *Instructional Technology: A Systematic Approach to Education*. New York: Holt, Rinehart, and Winston.
- Leshin, C.B., Pollock, J., and Riegeluth, C.M. (1992). *Instructional Design Strategies and Tactics*. Englewood Cliffs, New Jersey: Educational Technology Publications.
- Mager, R.F. (1962). *Preparing Objectives for Instruction* (2nd Ed.). Belmont, California: Fearon Publishers.

## Bibliography (Continued)

- Merrill, M.D., Tennyson, R.D., and Posey, L. (1992). *Instructional Design Strategies and Tactics*. Englewood Cliffs, New Jersey: Educational Technology Publications.
- Merrill, M.D., Lee, Z., and Jones, M.K. (1990). *Second Generation Instructional Design*. Englewood Cliffs, New Jersey: Educational Technology Publications.
- O'Neil, H.F., Jr., and Baker, E.L. (1991). Issues in Intelligent Computer-Assisted Instruction: Evaluation and Measurement. In T. Gutkin and S. Wise (Eds.), *The Computer and the Decision Making Process*. Hillsdale, New Jersey: Erlbaum Lawrence Associates.
- Reigeluth, C.M. (1983). Instructional Design; What is it and Why is it? In C.M. Reigeluth (Ed.), *Instructional Design Theories and Models? An Overview of Their Current Status*. Hillsdale, New Jersey: Erlbaum Associates.
- Rossett, A. (1987). *Training Needs Assessment*. Englewood Cliffs, New Jersey: Educational Technology Publications.
- Spears, W.D. (1983). *Processes of Skill Performance: A Foundation for the Design and Use of Training Equipment*. (NAVTRAEQ-VIPCEN 78-C-0113-4). Orlando, Florida: Naval Training Equipment Center.
- Tennyson, R.D., and Michaels, M. (1991). *Foundations of Educational Technology; Past, Present and Future*. Englewood Cliffs, New Jersey: Educational Technology Publications.
- Wolfe, P. et.al. (1991). *Job Task Analysis: Guide to Good Practice*. Englewood Cliffs, New Jersey: Educational Technology Publications.

### **Chapter 3**

## **GUIDELINES FOR CONSTRUCTING CRITERION-REFERENCED TESTS AND USING THE SURVEY TEST**

---

**Purpose of this chapter**

The information in this chapter is to be used in conjunction with the information contained in AFM 36-2234, Instructional System Development, and in AFH 36-2235, Information for Designers of Instructional Systems, Volumes 1-11.

The purposes of this chapter are to:

Provide guidelines for constructing Criterion-Referenced tests.  
Provide guidelines for using the survey test.

---

**Where to read about it**

This chapter contains three sections.

<b>Section</b>	<b>Title</b>	<b>Page</b>
A	Introduction	71
B	Guidelines for Constructing Criterion-Referenced Tests	73
C	Guidelines for Using the Survey Test	89

---

## Section A Introduction

---

### Introduction

Criterion-referenced tests are derived directly from the education or training objectives.

---

### Criterion-referenced tests and types of objectives

The following table describes the three types of Criterion-Referenced tests.

Type of Criterion-Referenced Test	Purpose	Developed from this type of objective
Criterion Test	To evaluate attainment of the objectives and to measure the effectiveness of the instructional system	Criterion Objectives
Diagnostic Test	To determine attainment of the supporting skills and knowledge for an instructional element.	Supporting Element Objectives <ul style="list-style-type: none"> <li>• Sub-objectives</li> </ul>
Survey Test	Administered during the analysis phase of instructional system development to determine what the target population already knows or can do before receiving instruction.	Supporting-Element Objectives <ul style="list-style-type: none"> <li>• Criterion Objectives (Terminal Objectives)</li> <li>• Sub-objectives (Enabling Objectives)</li> </ul>

---

### Developing the criterion-referenced test

The process of developing the Criterion-Referenced test involves four steps:

- Translating objectives into test items.
  - Developing the Criterion-Referenced test items.
  - Developing objective scoring procedures.
  - Trying out the Criterion-Referenced test.
-

**Characteristics of tests**

There are six basic characteristics that should be considered when developing tests to ensure they measure what is intended each time they are administered.

<b>Characteristic . . .</b>	<b>Refers to . . .</b>
Validity	The degree to which a test measures what it is supposed to measure.
Reliability	The degree to which a test yields the same results consistently.
Objectivity	The ability of a test to be free from variations due to factors other than the behavior being measured.
Comprehensiveness	The adequacy of a test to sample what is being measured.
Differentiation	The ability of a test to distinguish between levels of learning.
Usability	A test that is easy to administer, score, and interpret.

**Using the survey test**

Education and training requirements that are identified initially in the instructional system development process are based on “best guesses” about the abilities of potential students prior to exposure to the instructional system.

The survey test will either verify those guesses, or show where they are wrong.

A better instructional system can be designed if the designer knows:

If incoming students can already accomplish some of the proposed instructional objectives.

If some students can and some students can not accomplish the proposed instructional objectives.

If incoming students lack the prerequisite skills for entry into the proposed instructional system.

## Section B

### Guidelines for Constructing Criterion-referenced Tests

---

**Translating objectives into test items**

**Definition of a Valid Test Item**

A valid test item is derived from an objective that has been written to describe:

- The performance required.
  - The conditions of performance.
  - The standards required for speed and/or accuracy.
- 

**Examples of translating objectives into test items**

The following table describes examples of translation of two objectives into guidance for test development.

Objective	Test development guidance
Given a drawing of a skeleton, name 75 percent of the bones within 10 minutes.	Present a drawing of a skeleton with each bone numbered. Ask the student to write the correct name for each bone in the space next to the numbers. If the student correctly names 75 percent of the bones within ten minutes, the criterion for the objective is met.
Given an operable fire control radar system, measure (1) resistance, (2) ac voltage to $\pm 1000$ volts, (3) positive dc voltage to 1000 volts, (4) negative dc voltage to -1000 volts, and (5) positive dc current to 42.5 amps with a Multimeter. Record the values and determine if the measured value is within tolerance. Standard is at least 80 percent correct of each kind (1-5) of problem.	Ten problems of each type (1-5) will be given. The student must solve at least 80 percent of each type correctly to meet the criterion. The following test items are required: 10 resistors. Electrical power source that can provide 10 values and tolerances of the necessary voltages and current. Instructions for each problem. No hints will be provided. Scoring factors: (1) student writes measured value within tolerances, and (2) student uses correct procedure.

---

---

**Examples of translating objectives into test items (Continued)**

For the second example, observation of the procedure is important because the Multimeter has a variety of switch settings. Another reason for observing the procedure is to prevent damage to the Multimeter.

Although the second example is considerably more complex than the first example, the objective provides clear guidance for test development.

---

**General guidelines: Developing criterion-referenced test items**

Objectives must be prepared to Criterion-Referenced objective standards before tests are developed.

It's a good idea to develop all three types of tests (Criterion, Diagnostic, and Survey).

Criterion tests.

Survey tests substantiate education and training requirements. To be complete, survey tests must include *both* course criterion and diagnostic test items.

Diagnostic tests are required for identifying education or training problems.

---

**General guidelines: Translating objectives into test items**

To be sure that each objective has been properly translated into test items, compare each objective to the corresponding test item(s).

Identify, as specifically as possible, the *inputs to the student* (what the student is "given").

Identify the *correct student process and output*.

Ensure that the test items measure the learning behaviors and intellectual skills stated in the objectives, and that the performance and measurement standards of the test items are consistent with the objective standards.

Additional *inputs to the student* for each test item include a description of the test item (predictive or performance) that is appropriate for measuring the objective.

For a performance test item, note whether the problem involves a test of a product or process.

Specify the supplies and equipment needed for the test item.

---

---

**General guidelines:  
Translating  
objectives into test  
items  
(Continued)**

The *correct student process for responding* to a test item, and the *desired outputs* for each test item should be specified, including a description of how the test item is to be scored. For performance tests, note what part of the student's performance will be observed. Also note what will be considered an error.

---

**General guidelines:  
Decisions about  
predictive and  
performance test  
items**

Have decisions about predictive and performance test items reviewed by at least two subject matter experts. This ensures that the relationship between the objective and test item is as direct as possible. Ensure that:

The test item requires the student to produce the exact performance required by the objective, and no other.  
There are no ambiguous test item statements.  
The conditions under which the performance is to be observed are the same in the objective and the test item.  
The test item is scored according to the standard required by the objective.

---

**Guidelines for  
developing and  
reviewing test  
items**

The major problem in developing a test item is to clearly communicate the question or problem to the student. Test items should be developed using the following guidelines as a checklist. After the initial test items are developed, they should be reviewed again by a subject matter expert.

Keep the language simple. The ability of the student to comprehend difficult language ordinarily is not the skill in question.  
Tell the student whether speed or accuracy is more important, and whether there are any time limits for the test or a test item.  
Consider using graphics, photographs, video, audio, or other instructional media for test items, when appropriate for clear communication or for directly relating a test item to an objective.  
Present the test items so that they do not give the student hints related to the correct answer.  
Include any instructions common to all test items in the general overall test instructions.

---

**Guidelines for developing and reviewing test items**  
(Continued)

Provide clear instructions to the test administrator. Specify what should be said to the student, and how to answer student questions.

Arrange reasonable security to prevent students from receiving unplanned assistance or being disturbed while taking the test.

Give clear guidance to test administrators on when to excuse a student from a test, and under what conditions (such as equipment failure) scores may be considered invalid.

**Guidelines for developing predictive written test items**

A predictive test item presents a problem in the *item stem* that the student solves by *selecting* the correct answer from a set of alternatives or *producing* the correct answer from memory. The answer is referred to as the *item choice*.

For example, a test item problem could be posed as a sentence, a question, or a multiple choice selection:

“The types of test items requiring subjective scoring are \_\_\_\_\_.”

“What types of test items require subjective scoring?”

“The type of test item requiring subjective scoring is:

- a. multiple-choice
- b. true-false
- c. essay
- d. matching”

**Avoid true-false questions**

Try to avoid using true-false questions. They are the least reliable of the various types of predictive test items and generally are not suited for Criterion-Referenced testing.

**Guidelines for developing the test item stem**

The stem should *state the problem*. The student should know what the expected *test performance* is before reading the choices or making an answer production decision.

*Good Example*

Which of the following represents an example of a learning set?

*Poor Example*

Which of the following is best?

---

**Guidelines for the number of choices**

Allow for the number of choices you have decided upon.

*Good Example*

The female produces which of the following hormones?

- a. estrogen
- b. testosterone
- c. glucose
- d. gamma globulin

*Poor Example*

Which sex produces the hormone estrogen?

- a. female
  - b. male
  - c. both
  - d. neither
- 

**Guidelines for test item wording**

Word briefly and clearly.

*Good Example*

The type of test item most difficult to score is \_\_\_\_\_.

*Poor Example*

Test items may take many forms. The item most people have difficulty scoring is \_\_\_\_\_.

---

**Guidelines for test item grammar**

Use correct grammar within the item stem that is not relevant to the choices. Improper grammar may give away the choices.

*Good Example*

A mammal that is the only member in its zoological family is the:

- a. panda
- b. camel
- c. elephant
- d. koala bear

*Poor Example*

A mammal which is the only member in of its zoological family is an

- a. panda
  - b. camel
  - c. elephant
  - d. koala bear
-

**Guidelines for test item choices**

The following are guidelines for developing the test item choices:

Be clear and brief. Avoid wordiness. Be grammatically correct. Be parallel in content.

*Good Example (Clear, Brief, Grammatically Correct)*

The man who succeeded Lincoln to the Presidency was:

- a. James Buchanan
- b. Stephen Douglas
- c. Ulysses S. Grant
- d. Andrew Johnson

*Poor Example (Poor Content Parallelism)*

The man who succeeded Lincoln to the Presidency was:

- a. Christopher Columbus
- b. Dwight D. Eisenhower
- c. Napoleon
- d. Andrew Johnson

Construct incorrect choices that are plausible to students having varying degrees of information or misinformation. Incorrect choices can be derived from actual wrong answers given by students during instruction, from knowledge of common misconceptions, or from judgment of misconceptions students are likely to hold based on the instructional material. Plausible answers may also include phrases such as:

It cannot be determined from the information given.

None of the above.

All of the above.

(When using the above phrases, some test items should use the phrases as the correct answer).

Keep all alternatives relatively equal in length.

List choices containing numbers in descending or ascending order.

**Guidelines for test item choices (Continued)***Good Example*

On the Fahrenheit thermometer water freezes at:

- a. 0 degrees
- b. 10 degrees
- c. 32 degrees
- d. 64 degrees

*Poor Example*

On the Fahrenheit thermometer water freezes at:

- a. 10 degrees
- b. 64 degrees
- c. 32 degrees
- d. 0 degrees

**Guidelines for scoring predictive written test items**

In scoring items measuring intellectual skills, there may be acceptable variations in the wording of correct answers. Provide the test administrator with examples of correct and incorrect answers.

**Correcting for guessing**

In scoring objective predictive test items (such as multiple-choice), consider correcting the test for guessing. You should correct for guessing when:

A student can identify the correct response in a test item either by knowing it or by guessing it.

Wrong answers are obtained exclusively by guessing. A wrong response means lack of knowledge and consequent guessing.

A wrong answer may have been selected because the student thought it was the correct answer and not because the student guessed. In this case, the student has not acquired the correct information during instruction (lack of knowledge).

**Formula for adjusting scores for guessing**

The formula for adjusting scores for guessing is:

$$\text{Corrected Score} = R - \frac{W}{K - 1}$$

Where:

R = number of items answered correctly

W = number of items answered incorrectly

K = number of alternatives to each item

**Example of correction for guessing**

A 100-item multiple-choice test has four alternatives to each item. The number of right and wrong responses for three students is shown in the following table.

	Number Right	Number Wrong	Number Attempted	Corrected Scores
<b>Student A</b>	79	21	100	72
<b>Student B</b>	76	9	85	73
<b>Student C</b>	74	0	74	74

When corrected for guessing, Student C, who actually knew the correct responses to more items, gets the highest score.

---

**Guidelines for developing performance test items**

The performance test item requires the student to perform a task or some portion of a task.

First, it is important to determine whether a *process* or a *product* is being measured by the performance test.

Next, determine if the checklist, numerical, descriptive, or graphic rating scale is best to use.

As with predictive written test items, be sure that the conditions of performance, the performance behavior, and the criterion standard in the checklist or other rating scale parallel the conditions of performance, the performance behavior, and the criterion standard in the education/training objective.

---

**Performance tests with supplies and equipment**

General guidelines for developing performance test items include:

Prepare a complete, accurate list of any supplies needed for the test.

When performance tests involving use of equipment (e.g., a computer, training device, or training materials) are required at specific test stations, plans for the stations should include:

The complete list of equipment and supplies needed.

---

**Performance tests with supplies and equipment (Continued)**

How the station is to be set up for each student. This includes station layout, equipment switch settings, and detailed descriptions of equipment operation for the test. Directions for the test administrator to prevent injury to the student or damage to the equipment.

Directions for the test administrator explaining the procedures for setting up the station again before the next student arrives.

---

---

**Guidelines for  
scoring  
performance test  
items**

Scoring procedures are often a major source of unreliability or inconsistency in performance tests. Scoring procedures should be developed with care.

**Scoring Objectives**

The test administrator should not rely on judgment or general impressions. Depending on the performance objective, either a *product* or a *process* may be rated or measured.

Several trained scorers should be able to score a given student and arrive at the same score on a given test or item. Item objectives should state a time and/or accuracy standard that must be met to meet the criterion for performance.

The objective standards are the basic guides for developing scoring procedures.

Strive for complete objectivity in scoring.

**Developing Scoring Procedures**

To develop scoring procedures, first study the objective the item is to measure.

If the criterion standard is one of accuracy, determine specifically what in the student's performance should be observed or measured.

Clearly identify observable characteristics so that the test administrator (scorer) can make a record of observations for each characteristic.

---

**Product  
measurement**

In general, products are measured under the following conditions, using a rating scale:

**When to measure**

When the performance objective clearly calls for the student to produce something.

---

---

**Product  
measurement  
(Continued)**

When the product can be readily evaluated. This implies that the product has two important features:

It can be judged by significant features that are present or not present.

It can be measured accurately by instruments to determine physical characteristics such as weight, conformity to specifications, proper electric voltage/current, wave forms, etc.

When the student does not have to follow a definite, fixed, procedure.

If student performance is less important than the production of a satisfactory product

**How to measure**

In developing a scoring procedure for a product, identify as accurately and specifically as possible the characteristics of the product to be scored.

Define the specific characteristics that distinguish a satisfactory product from an unsatisfactory product.

If the product is to be measured by some kind of an instrument, identify the characteristics to be measured.

Provide the test administrator with specific instructions for making the measurement.

Define the criterion (passing) score for each performance test item based on the time and/or accuracy standards stated in the performance objective.

If an objective does not provide clear standards, revise the objective until it does.

---

**Process  
measurement**

Process measurement is appropriate using a checklist or another form of a rating scale when:

The performance objective calls for correct performance of a *sequence of actions*, or for the *cognitive generation of strategies or tactics*.

The way a task is performed is as important as the final product.

The sequence of correct actions can be observed.

---

---

**Process measurement: Test administrator directions**

To get a good scoring procedure for process measurement, provide the test administrator with explicit directions on what the student should be doing at each stage in the task.

Provide a step by step description of how the task is performed.

Place the explicit step performance actions in checklist form.

---

**Guidelines for trying out the test**

There must be a test tryout. Each of the many different forms that tests can take has problems peculiar to the form. The tryout of a draft test will identify and correct sources of unreliability in the test items.

---

**Test tryout: Correcting for unreliability**

To identify and correct unreliability in test items, the developer must:

Conduct the draft test.

Draft general instructions for the test.

Train test administrators and raters. Use at least two, and preferably three scorers.

Select students. Select at least 20 who are typical of the students in the course. If selection of 20 students is not possible, use as many as possible.

Prepare forms for recording information concerning each item. Select observers. Observers watch the administration of the test and record information that can be used to correct deficiencies in the test items or in the test procedures. Observers can also be used as scorers.

---

**Test tryout: Make it real**

As a general rule, conduct the tryout as if it were "for real."

Conduct the tryout in a sequential fashion.

Test five students, and revise the test to correct any difficulties found. Then test the next five students as a check of the revisions.

Check the success (or lack of success) of the revisions.

Continue the process of testing and revisions until all deficiencies are corrected.

---

---

**Test tryout: Make it real (Continued)**

Record information in sufficient detail to provide a basis for correcting any deficiencies in test items. The observer should use the following techniques for recording information concerning test items:

**The Test Environment**

Record any shortage of supplies or breakdown of equipment. Note any ways in which the layout of equipment can be improved without impairing the validity of the test.

Note any accidental injury to the student or damage to the equipment.

Note the time required and any problems encountered in reconfiguration of the student test station for the next student.

If the test is given in a series of test stations, record any problems experienced in maintaining a smooth flow of students from station to station.

**The Students**

To see if general and specific instructions to the student are clearly understood, ask the student to repeat them verbally.

Note any significant deviations.

Record any questions asked by students. Prepare written instructions to cover points on which questions are often raised.

Note any conditions that may render a test item invalid.

Question each student whenever an error is observed.

Determine if there is a misunderstanding of the test item.

Note any actions of the test administrator that might give away the correct answer to a test item, or that might confuse the student.

---

**Test tryout: Correct for deficiencies**

The goal in developing a scoring system is to have written scoring instructions that yield the same score, regardless of who does the scoring. The test tryout will determine how well the scoring system works. The following guidelines should be applied during test development:

Pose clear, specific problems or tests to the student.

Score only performance or product for characteristics that are observable.

Apply a definite standard to determine whether the student passes or fails.

---

---

**Test tryout: Correct  
for deficiencies  
(Continued)**

Avoid scoring procedures that call for the scorer's judgment or opinion, rather than observation of facts.

---

**Guidelines for  
developing a  
scoring system**

Have each item scored independently by at least two people.

The scorers should not communicate, compare notes, or compare scores until all scores have been recorded.

Note the score that each rater has given to each student. If the scores are different, there are several possible reasons:

The scorers did not observe the same characteristics of performance or product.

The form on which observations were made was inadequate.  
The scorers need more training.

---

**Guidelines for  
scoring during test  
tryout**

Some point to consider are:

The scorers observed the same thing, but did not apply the same standard.

The scorers did not agree on a standard consistent with the objectives.

The scorers were overloaded, and could not observe all required items. This is most likely to happen when using a checklist to rate a process.

If certain critical portions of the process can be selected, the number of required observations can be reduced.

If certain critical portions of the process can not be selected, conduct additional training for the observers.

The scorers were inconsistent in applying measuring techniques.

For example, if a characteristic of a product must be measured by a test instrument, scorers could use the instrument in different ways.

Establish consistent measuring procedures.

---

---

**Purpose of the test tryout**

The purpose of the test tryout is to make the test as reliable as possible by eliminating as many sources of unreliability as possible. Some of the do's and don'ts of the test tryout are:

Don't use the test tryout to grade students.

Don't use the test tryout to test each item on the test to the same extent.

Drop test items from the tryout after they have been judged acceptable following administration to at least 10 students.

Be aware of test items that are closely related or that are dependent upon each other. In such cases, retain the set of test items until all of the items are judged to be satisfactory.

---

**Guidelines for using a test construction worksheet**

Figure 1 depicts a notional test construction worksheet for the development of a *product* test item.

Figure 2 depicts a notional test construction worksheet for a *performance* test item for a flying maneuver.

The worksheets are not appropriate for recording most predictive written test items.

---

Figure 1 Notional Test Construction Worksheet for Product Test Items

Education/Training Objective Number: \_\_\_\_\_

<p><b>Input to Student</b></p>	
<p><b>Instructions</b></p> <p><b>Questions</b></p> <p><b>Item Stem Including Alternatives</b></p> <p><b>Problem Aids</b></p>	
<p><b>Correct output</b></p>	
<p><b>Answer</b></p> <p><b>Product</b></p> <p><b>Performance Including Rating Scales</b></p>	



## Section C

### Guidelines for Using the Survey Test

---

#### **Lessons learned for using the survey test**

The following are lessons learned for use of the survey test:

Use the survey test to assess entering behavior.

Administer the survey test to a sample of the target population.

Use the survey test results as feedback to:

Determine the adequacy of the education/training requirements in the proposed course of instruction.

Determine the relevancy of the criterion objectives in the course of instruction in relation to any changes in the education/training requirements.

Determine the relevancy of the test items in the course of instruction in relation to the criterion objectives and to any changes in the education/training requirements or the criterion objectives based on the survey test results.

---

#### **Constraints when using the survey test**

The following constraints could occur when using the survey test:

Performance of job tasks by untrained persons may result in danger for personnel or damage to equipment.

Expensive equipment or training devices may be needed to assess the capabilities of students entering the course of instruction.

The target population that is representative of students entering the course of instruction may be widely dispersed or unavailable for sampling.

Facilities for conducting the survey test may be inadequate or unavailable.

Limited time may be available for administration of the survey test.

Limited funds may be available for administration of the survey test.

---

---

**Guidelines for  
using the survey  
test**

Guidelines for administration of the survey test include the following:

**Preparation and Testing**

Make sure that the sample of students is representative of the target population.

Plan to administer the survey test to at least 10 students.

If the test is extremely long, students need not take the entire test. However, at least 10 students should be tested on every test item.

Make the test situation as realistic as possible.

Include items testing all criterion objectives.

Tell the students that you are using the test to develop a new instructional program.

Encourage the students to do as well as they can.

**Scoring and Test Revision**

After administering the survey test, score it as reliably as possible.

Results of the survey test can indicate revision of objectives and test items.

If the survey test shows that an objective is inappropriate, revise the objective and related test items.

Review, and revise as necessary, the education/training requirements associated with the objective.

If the results of the survey test require substantial revisions to the criterion objectives, a revised survey test will have to be developed and administered.

---

**Guidelines for actions based on the results of administration of a survey test**

The following table describes results that could occur following administration of a survey test, and actions that should be taken in response to the results.

<b>Results of Survey Test</b>	<b>Action to Take</b>
All students can exhibit a particular criterion behavior.	Do not prepare instruction for that criterion behavior.
Some students can exhibit a criterion behavior, and some cannot.	Consider allowing more able students to bypass the related instructional material.
A majority of the students experience greater difficulty with the criterion behaviors than anticipated.	Review the diagnostic test results to diagnose the source of difficulty, and increase instructional content in these areas, as necessary. (The diagnostic test items are part of the survey test.)
A majority of the students experience less difficulty with the criterion behaviors than anticipated.	Review terminal and enabling objectives to determine how the amount of instructional content can be reduced. Delete education/training requirements as required.

## Bibliography

- Bills, C.G., and Butterbrodt, V.L. (1992). *Total Training Systems Design Function: A Total Quality Management Application*. Wright-Patterson AFB, Ohio.
- Briggs, L.J., and Wager, W.W. (1981). *Handbook of Procedures for Design of Instruction* (2nd Ed.). Glenview, Illinois: Harper Collins Publishers
- Carlisle, K.E. (1986). *Analyzing Jobs and Tasks*. Englewood Cliffs, New Jersey: Educational Technology Publications.
- Davies, I.K. (1976). *Objectives in Curriculum Design*. London: Mc Graw Hill.
- Dick, W., and Carey, L. (1990). *The Systematic Design of Instruction* (3rd Ed.). Glenview, Illinois: Harper Collins Publishers.
- Gagné, R.M. (1985). *The Conditions of Learning* (4th Ed.). New York: Holt, Rinehart and Winston.
- Gagné, R.M., Briggs, L.J., and Wager, W.W. (1992). *Principles of Instruction* (4th Ed.). New York: Harcourt Brace Jovanovitch College Publishers.
- Gagné, R.M., and Merrill, M.D. (1990). *Integrative Goals for Instructional Design*. Englewood Cliffs, New Jersey: Educational Technology Publications. 38(1), 1-8.
- Goldstein, I.L. (1986). *Training In Organizations: Needs Assessment, Development, and Evaluation* (2nd Ed.). Pacific Grove, California. Brooks/Cole Publishing Company.
- Hageman, D.C. (1988). *Cognitive Engineering of Training Systems for Simulators*. National Aerospace and Electronics Conference. Dayton, Ohio.

## Bibliography (Continued)

- Hageman, D.C. (1985). *Effective Training Systems for High-Technology Equipment Operation*. National Security Industrial Association Fifth Annual Conference on Personnel and Training System Effectiveness. San Antonio, Texas.
- Keller, J.M. (1987). The Systematic Process of Motivational Design." *Performance and Instruction*, 26(9), 1-8.
- Kibler, R.J. (1981). *Objectives for Instruction*. Boston: Allyn and Bacon.
- Knirk, F.G., and Gustafson, K.L. (1986). *Instructional Technology: A Systematic Approach to Education*. New York: Holt, Rinehart, and Winston.
- Leshin, C.B., Pollock, J., and Riegeluth, C.M. (1992). *Instructional Design Strategies and Tactics*. Englewood Cliffs, New Jersey: Educational Technology Publications.
- Mager, R.F. (1962). *Preparing Objectives for Instruction* (2nd Ed.). Belmont, California: Fearon Publishers.
- Merrill, M.D., Tennyson, R.D., and Posey, L. (1992). *Instructional Design Strategies and Tactics*. Englewood Cliffs, New Jersey: Educational Technology Publications.
- Merrill, M.D., Lee, Z., and Jones, M.K. (1990). *Second Generation Instructional Design*. Englewood Cliffs, New Jersey: Educational Technology Publications.
- O'Neil, H.F., Jr., and Baker, E.L. (1991). Issues in Intelligent Computer-Assisted Instruction: Evaluation and Measurement. In T. Gutkin and S. Wise (Eds.), *The Computer and the Decision Making Process*. Hillsdale, New Jersey: Erlbaum Lawrence Associates.
- Reigeluth, C.M. (1983). Instructional design; What is it and Why is it? In C.M. Reigeluth (Ed.), *Instructional Design Theories and Models? An Overview of Their Current Status*. Hillsdale, New Jersey: Erlbaum Associates.
- Rossett, A. (1987). *Training Needs Assessment*. Englewood Cliffs, New Jersey: Educational Technology Publications.
- Spears, W.D. (1983). *Processes of Skill Performance: A Foundation for the Design and Use of Training Equipment*. (NAVTRAEQ-VIPCEN 78-C-0113-4). Orlando, Florida: Naval Training Equipment Center.
- Tennyson, R.D., and Michaels, M. (1991). *Foundations of Educational Technology; Past, Present and Future*. Englewood Cliffs, New Jersey: Educational Technology Publications.
- Wolfe, P. et.al. (1991). *Job Task Analysis: Guide to Good Practice*. Englewood Cliffs, New Jersey: Educational Technology Publications.

## Chapter 4

### GUIDELINES FOR VALIDATION OF INSTRUCTIONAL RESOURCES

---

**Purpose of this chapter**

The information in this chapter is to be used in conjunction with the information contained in AFM 36-2234, Instructional System Development, and in AFH 36-2235, Information for Designers of Instructional Systems, Volumes 1-11.

The purposes of this chapter are to:

Provide guidelines for validating instructional resource requirements, including equipment, training devices, ICW, audiovisual media, hardcopy media, facilities, manpower, and costs.

Provide guidelines for analyzing tryout materials and making revisions.

Provide guidelines for conducting a validation of the instructional system.

---

**Where to read about it**

This chapter contains six sections.

Section	Title	Page
A	Introduction	94
B	Validation of Resource Requirements	95
C	Validation of the Instructor	100
D	Validation of Instructional Materials	105
E	Analyzing Tryout Materials and Making Revision	109
F	Conducting a Validation of the Instructional System	125

---

## Section A Introduction

---

### Validation of instructional resources

Once the sequence of instruction, the instructional strategy, the instructional methods, and the instructional media have been selected, and the initial instructional materials have been developed, it is time to validate the instructional resources.

Validation of instructional resources includes:

- Validation of resource requirements.

- Validation of the instructors.

- Validation of the instructional materials, *including tests and test items*.

- Tryout of the instructional materials and making appropriate revisions.

- Conducting a validation of the operation of the instructional system.

---

## Section B

### Validation of Resource Requirements

---

#### Introduction

Early in the development stage of an instructional system, it is important to identify and validate instructional resources such as equipment, training devices, interactive computer-based courseware, audiovisual media, hardcopy media, facilities, manpower needs, and costs.

If the instructional system development process is followed correctly, instructional system resource requirements will be determined and validated before instructional materials are developed.

---

#### Validation: Resources and cost factors

The following table summarizes validation considerations for instructional system resource requirements and cost factors.

Resources	Cost Factors
<b>Equipment</b> Instructional Equipment Support Requirements  <b>Training Devices</b> Instructional Devices Support Requirements	<b>Development Costs</b> Equipment Training Devices Facilities Materials
<b>Interactive Computer-based Courseware (ICW)</b> Instructional ICW Support Requirements	<b>Investment Costs</b> Equipment Training Devices Facilities Materials
<b>Audiovisual Media</b> Instructional Media Support Requirements  <b>Hardcopy Media</b>	<b>Operation and Maintenance Costs</b> Equipment Training Devices Facilities Materials
<b>Facilities</b> Academic Classrooms Laboratories Training Device Facilities Special-Purpose Facilities	Pay and Allowances

---



---

**Validation:  
Resources and  
cost factors  
(Continued)**

Resources	Cost Factors
<b>Manpower</b> Instructors Administrators Supervisors Base Administration and Support	

**Equipment**

Instructional equipment includes any item or combination of items used for instruction. For example, computers and associated equipment, and audiovisual media hardware such as projectors, screens, cameras, training aids, slides, transparencies, etc.

Support equipment includes items such as computer software, chairs, desks, computers, filing cabinets, power units, etc.

**Training devices**

Training devices include items or combination of items used for instruction, for example, aircraft, part-task training devices, simulators, etc.

Support equipment includes all spare parts and equipment required to operate and maintain the training devices.

**Interactive  
Computer-based  
Courseware (ICW)**

ICW includes items or combination of items used for instruction, for example, authoring systems, digitized photograph or video files, digitized audio files, animated text and graphic files, educational/training requirements analysis documents, media analysis documents, storyboard documents, etc.

Support equipment includes all software and hardware for the development or revision of ICW, and the spare parts, equipment, and data files required to operate and maintain the ICW.

**Audiovisual media**

Audiovisual media includes items or combination of items used for instruction, for example, slides, view graphs, videotapes, etc.

Support equipment includes all equipment used to accomplish revision requirements, spare media, and the equipment and spare parts required to display the media.

---

**Hardcopy media**

Hardcopy media includes items or combination of items used for instruction, for example, student and instructor guides, workbooks, reference materials, technical documents, manuals, regulations, etc.

Support equipment includes all equipment and spare parts used to accomplish revision requirements, spare media, and storage requirements.

---

**Facilities**

An instructional facility is the physical complex in which instruction is conducted, and the physical areas which provide direct support for instruction. Some examples include academic classrooms, laboratories, training device facilities, and special-purpose facilities.

---

**Manpower**

Manpower includes all personnel required to accomplish the instructional system mission and associated workloads. Manpower requirements for an instructional system include instructors, administrators, administrative personnel, education/training specialists, computer programmers, courseware authors, etc.

---

**Costs**

Certain costs are associated with each required instructional system resource. In determining costs for an instructional system, consider three types of costs:

**Cost of Acquiring and Developing the Resources**

This includes acquisition costs for instructional devices and equipment, support equipment, facilities, the design and development of course materials, training instructors, and any special equipment or facilities. Development costs include pay and allowances.

**The Investment the Costs Represent**

This includes the per-student costs and the potential for future use and reuse of instructional system resources.

**The Costs of Operation and Maintenance**

This includes the costs after acquisition of the instructional system to keep training devices, equipment, courseware, and facilities up to standard. It also includes the costs of replacing expendable or consumed materials, and the costs of revising instructional materials.

---

---

**General validation considerations for instructional resource selection**

Validation of selected instructional resources will be influenced by the following considerations:

The level of the criterion standards for the course of instruction.

Student entry rate and group size for the course of instruction.

Time limits for the conduct of the course of instruction and for the students to attain the criterion standards for the course of instruction.

The planned use of the instructional resources in the course of instruction.

---

**Validation considerations for training devices and equipment**

Points to consider:

Is duplication of training devices or equipment necessary because a large number of students will be attending the course of instruction? For example, must several students or several groups of students receive demonstration, practice, and evaluation of performance tasks simultaneously?

Can performance instruction be staggered to make fewer training devices or equipment available to more students over a period of time?

Do facility limitations exist that will affect training device or equipment considerations? For example, is adequate power, floor space, an area for briefing/debriefing/classroom use, environmental control equipment, etc., available to support the training devices and equipment?

What personnel capabilities are needed? For example, do instructors need special training to use the training devices or equipment for instruction? Do the training devices or equipment require expenditure of funds for maintenance or supply of spare parts?

Are suitable "off-the-shelf" commercial or military training devices or equipment available?

Can existing commercial or military training devices or equipment be modified economically to meet the education/training requirements?

Can existing commercial or military training devices or equipment be cross-utilized with other instructional systems without impairing the mission of another instructional system?

---

**Validation considerations for training devices and equipment (Continued)**

Is the size and configuration of the classrooms, briefing/debriefing areas, laboratories, or other special-purpose areas associated with the training devices or equipment adequate? Are the facilities adequate for the required number of students, instructor personnel, and instructional materials or equipment?  
Does the facility meet requirements for environmental controls, light, acoustics, etc.?

---

**Validation considerations for facilities**

Points to consider:

What are the Air Force regulations and criteria regarding instructor personnel selection?  
What is the student entry rate and group size?  
What is the appropriate instructor/student ratio?  
Do instructor personnel need special training or qualifications because  
the tasks to be taught are complex, involve safety hazards, or require special equipment? Do other existing instructional systems have instructor personnel available for reallocation?

---

## **Section C**

### **Validation of the Instructor**

---

**Introduction**

The instructor is a vital part of the validation of any instructional system, and is the key to successful implementation. Key activities that an instructor must perform include:

- Conducting the instruction as it was designed.
  - Ensuring that the students are participating actively in the course.
  - Assessing and analyzing student performance.
- 

**Validation considerations for instructor orientation**

The following items should be considered in order to orient the instructor to the instructional system:

- Is the instructor prepared to undertake the functions required by the instructional system? Individualized, self-paced instructional systems require the instructor to facilitate learning in a manner that may differ from an instructor's past experience in classroom instruction.
  - Has the instructor been involved in the instructional system development and validation process to ensure that the course is implemented as designed?
  - Can the instructor accept new or different instructional methods or philosophies? Assist the instructors by providing them with education/training on the application of new concepts and techniques, as well as the rationale for their use. Show instructors how they will benefit from new instructional approaches.
- 

**Validation considerations for the primary roles of the instructor**

The instructor has three primary roles. The instructor must accomplish the tasks required for a course administrator, facilitate delivery of the instruction, and perform the tasks required of an individual tutor or counselor.

---

---

**Role of a course administrator**

In the role of a course administrator, the instructor must be able to:

Ensure that the instructional materials, equipment, instructional aids, and other supplies are readily accessible to the students?

Ensure that training devices, equipment, and instructional aids function properly?

Ensure that each student is progressing within the planned scope of the teaching-learning activities?

Gather data on the instructional segments where recurring deficiencies in student performance occur? These data are important for revisions to an instructional segment.

Administer tests associated with the instructional system to determine student achievement of the course objectives?

Gather and record all performance data for both the students and the instructional system?

Monitor student progress in the instructional system, and coordinate the awarding of reinforcers?

When the instructional segment is modular or self-paced, can the instructor schedule students, training devices, and equipment, for effective instructional activity in accordance with the course schedule and resource availability?

---

**Role of an individual tutor or counselor**

In the role of an individual tutor or counselor, is the instructor able to:

Observe the failure of students to meet objective criteria, to understand teaching points, or to perform certain tasks?

Provide assistance to students to help them resolve their difficulties?

Rechannel student activities into remedial instructional segments or provide individual tutoring as required?

Observe unfavorable student attitudes, opinions, or emotions that can reduce student ability to learn?

Interact individually with students to help them overcome unfavorable attitudes, opinions, or emotions before they seriously affect learning?

Understand the course objectives and student capabilities in order to know when and how to provide assistance?

---

---

**Validation considerations for instructor preparation**

Instructors may not be accustomed to Criterion-Referenced instruction that emphasizes how the student performs, rather than how the instructor performs. Have the instructors been prepared for their roles by:

- Learning about new techniques and technologies?
- Attending orientation programs that have provided interaction with other instructors?
- Participating in the planning and development of new instructional techniques and technology?
- Involvement in the exchange of ideas concerning the improvement in instructional conditions and resources, instructional technologies, methods of presentation, and evaluation procedures?
- Interchanging ideas with other instructors who are instructing in other courses at the same base or other bases?
- Observing new instructional techniques or technologies in action?
- Providing inputs to the Instructor Guide for the course of instruction that includes directions for carrying out the course? (The Instructor Manual should be an integral part of the course design and development phase of the instructional system development process.)

---

**Validation considerations for the instructor guide**

The Instructor Guide should include sections that provide the instructor with the

- Course Description
- Target Population Description
- Criterion Tests
- System Performance Data
- Directions for Administering the Course

The following tables describe recommendations for validating the contents of these five sections

---

## Validation Considerations for the Instructor Guide

<b>Validation: Course Description</b>
<p>Brief statement of the purpose and scope of the course, including the student competencies upon completion of the course, what job the student will be prepared to perform, and a description of the student population for which the course is intended.</p> <p>An overview of the contents of each instructional Module, Unit, Block, Lesson, and Lesson Segment.</p> <p>A syllabus or plan of instruction in the proper learning sequence for each instructional Module, Unit, Block, and Lesson.</p> <p>Lesson Plans for each lesson in the course arranged by instructional sequence by type of instructional media. Lesson Plans should include</p> <ul style="list-style-type: none"> <li>associated terminal and enabling objectives.</li> <li>the instructor presentation requirements.</li> <li>the expected student activities.</li> <li>the instructional media required and how to use the media to administer the lesson. (For example, an aircraft, a simulator, a part-task training device, an interactive courseware learning laboratory, a self-paced learning environment, or an academic classroom)</li> </ul>
<b>Validation: Target Population Description</b>
<p>Educational level of the student population.</p> <p>Previous training and related knowledge of the student population.</p> <p>Required physical and personal characteristics of the student population.</p>
<b>Validation: Criterion Tests</b>
<p>Answers for tests or measurement instruments on the instruments themselves or on a separate answer sheet.</p> <p>Directions for administering tests and measurement instruments.</p> <p>Rating, scoring, and weighting procedures.</p> <p>Reference to criterion objectives tested or measured by individual test items.</p>
<b>Validation: System Performance Data (historical)</b>
<p>Description of the validation and evaluation processes used for the course of instruction.</p> <p>Comparative pre-test and post-test results and other historical system performance data.</p>

**Validation Considerations for the Instructor Guide (Continued)****Validation: Directions for Administering the Course**

Directions for orientating the students to the training situation.

Instructional material required by the students.

Directions for conducting the instruction, including:

Scheduling procedures.

Procedures for handling individual student differences.

Processes for monitoring instruction.

Procedures for keeping students productively involved in the learning process.

Recommendations for handling exceptionally fast or unusually slow students.

Recommendations for providing an environment conducive to learning.

---

## Section D

### Validation of the Instructional Materials

---

#### Overview of the internal review of the instructional material (formative evaluation)

Draft instructional material should be subjected to an internal review by instructional system development personnel before a small-group tryout of the instructional material is conducted.

*The **internal review** consists of individual, single-group, and small-group tryouts. The internal review is a **formative evaluation** activity, and is the first step of the actual validation process for instructional materials.*

---

#### Operational tryout (summative evaluation)

Following the internal review (formative evaluation) process, the instructional materials are subjected to an **operational tryout**. The operational tryout is the final step in the validation process and is the **summative evaluation** of the instructional materials.

---

#### Operational evaluation of the instructional system

Following the summative evaluation of the instructional materials, an **operational evaluation of the instructional system** is conducted. Operational evaluation consists of **internal evaluation** and **external evaluation**. Chapter 5 describes the operational evaluation process.

---

#### Conducting an internal review of instructional materials

Instructional materials should not be operationally tried out on students until an internal review (formative evaluation) has been conducted using individual and single-group tryouts. Draft instructional material is likely to contain technical inaccuracies and flaws that can be identified and corrected before a small-group tryout is conducted.

---

#### Internal review: Personnel characteristics

Personnel assigned to conduct an internal review of instructional materials should have these characteristics:

- Have broad knowledge of the process and techniques of instructional system design.

- Be expert in the content area of the instructional material.

- Have demonstrated the ability to be a concise and constructive critic.

---

---

**Internal review:  
Instructional  
materials**

Submit instructional materials for internal review as they are developed. Examples of the types of instructional materials to submit are:

Terminal and enabling objectives.  
Test and measurement items.  
Draft instructional materials and Instructional System Development documents.

---

**Internal review:  
Review questions**

Some questions that the reviewer should ask during internal review of the draft instructional materials are:

Is the content accurate?  
Which are the “good” and “bad” sections of the instructional system and what are the reasons for this judgment?  
Is the instructional material sequenced effectively?  
Are the practice, remediation, and review opportunities adequate?  
How effectively does the instructional material teach the specific behaviors specified in the terminal and enabling Criterion-Referenced objectives?  
Do the criterion referenced tests and test items directly test and measure the specific behaviors specified in the terminal and enabling Criterion-Referenced objectives?  
Are there test and measurement items for each terminal and enabling objective in the course of instruction?

---

**Internal review:  
Detailed comments**

The reviewer should make detailed comments during internal review of the instructional materials. Documentation of problems in the instructional materials should be very specific, and should contain recommendations for correction or modification.

---

**Internal review:  
Discussing results**

Once the internal review has been completed, the instructional system development personnel should discuss the results with the reviewers. Some considerations are:

Comments by a reviewer should be considered as suggestions only.

---

---

**Internal review:  
Discussing results  
(Continued)**

Different approaches to correcting a problem may be equally valid.

The results of the internal review do not provide conclusive proof of the adequacy or inadequacy of the instructional materials.

The results of the internal review do identify some of the potential problem areas, and provide suggestions where improvements may be made.

---

**Conducting an  
individual, single-  
group, and small-  
group tryout of  
instructional  
materials**

As part of the internal review (formative evaluation) of the instructional material, a tryout of the draft instructional materials is conducted on an individual, single-group, and a small-group of individuals from the target population of potential students. The following are guidelines for conducting an internal review of instructional material:

First, tryout the instructional materials on individual students.

Use a single-group of students if the instructional materials require that several individuals must work together.

After the instructional materials have been tried on two to five students or a single-group of students, revise the elements of the instructional material where the students had difficulty.

Next, try the revised instructional material on an additional two to five students or student groups.

After the instructional materials have been tried again on two to five students or single student groups, revise the elements of the instructional material where the students had difficulty.

Continue this process as long as improvement to the instructional materials is required, and time and money permit.

---

**Content revisions**

The content of the instructional materials, including the test and measurement instruments, should not be revised on the basis of a single student's errors, unless technical inaccuracies or obvious deficiencies are discovered.

If technical revisions are made to the instructional materials, revise the associated terminal and enabling objectives as well as the test and measurement materials.

At least two to five students should try out the instructional materials before revisions are made.

Look for consistent trouble spots and errors on the part of several students.

---

---

**Importance of the tryouts**

The importance of these instructional material tryouts cannot be minimized. Accomplish the following actions if students are not available for instructional material tryouts and internal review of the instructional materials:

Attempt to find personnel who have qualifications similar to those of the target population to use for the individual tryout. If these personnel are not available, carefully analyze the first students who attempt the instruction during the administration of the instructional materials to an appropriate sample of the actual student population.

---

**Selecting an appropriate sample for individual, single-group or small-group tryouts**

The following are guidelines for following student selection guidelines for the individual, single-group, or small-group tryouts:

Select students who fall within the range of aptitudes, prior knowledge, skills, background, and attitudes of typical students for the course of instruction. If the sample of students does not fall within the range of typical students, the results of the small-group tryout will be biased. The results of the small-group tryout will not be able to be generalized to the actual target population of students. The sample students used in the tryouts should not represent "average" students. They should come from the upper 25% in aptitude and background. The reason for selecting these students for the tryout are:

More capable students often can help point out and analyze weak spots in the instruction.

If more capable students cannot learn from the material and master the Criterion-Referenced objectives, the less capable student certainly cannot.

If less capable students are administered the test and measurement instruments before the more capable students, there is no way to tell if the instructional material is too basic, if there are too many teaching points, or if there is too much practice time in the proposed syllabus.

It is easier to revise the instructional material from a known point of difficulty with the more capable students down to the learning level of the less capable students.

---

---

**Selecting an appropriate sample for individual, single-group or small-group tryouts  
(Continued)**

It is difficult to revise the instructional material from an unknown point of difficulty with the less capable students up to the known point of difficulty of the more capable students.

It is simpler to add material to make instructional materials easier to master than it is to delete material to make instructional materials more difficult to master.

---

**Administering the instructional materials for individual, single-group, or small-group tryouts**

The following are guidelines for administration of the instructional materials to an individual, single-group, or small-group:

Use the identical media in the tryouts that were selected for use in the instructional system.

Prepare the students for the tryouts. Inform the students that they are not being evaluated during the tryouts. Inform them that they are evaluating the instructional materials.

Use these sources of information during the tryout:

*Diagnostic Tests:* Administer a diagnostic test as a pre-test to identify the entering capabilities of the students. Administer the diagnostic test as a post-test to assess learning due to the instruction. The post-test will identify errors in the instructional material and weak points in the instructional system.

*Observation of Student Performance:* Observe and record student performance during exposure to the instruction. Obtain information about which exercises, tasks, objectives, lessons, teaching points, etc., result in student errors. Observe what type of errors are being made. Observe how many students make a specific error.

*Student Comments:* After students complete the post-test, gather their reactions about any difficulties they encountered during instruction. Ask students for suggestions on how to improve the instruction. Devise a questionnaire to get systematic answers and comments.

---

---

**The role of the instructor during individual, single-group, or small-group tryouts**

The instructor determines the adequacy of presentations and supporting instructional media through feedback from the students.

During the tryout, the instructor should note any problem areas.

The instructor should refrain from providing additional assistance to the student unless it is absolutely necessary for the student's progression in the course of instruction.

After the student completes the instruction and the associated test and measurement instruments, the instructor should encourage the student to discuss any difficult areas encountered during instruction.

---

## **Section E**

### **Analyzing Tryout Test and Measurement Items and Making Revisions to Instructional Materials**

---

#### **Introduction**

Each test and measurement item contained in the instructional materials should be based solely on the requirements specified in the objectives that the test and measurement item is to measure.

If a student fails the test items associated with an objective, the objective has not been mastered.

Normally, at least 80% of the students should pass a Criterion-Referenced test item.

If a large percentage (e.g., 80%) of the students do not pass a Criterion-Referenced test item, the individual, single-group, or small-group internal review (formative evaluation) data must be analyzed to determine why the students are having difficulty.

Revisions must be made to either the design of the instructional system or to the instructional materials if 80% of the students do not pass Criterion-Referenced test items.

---

#### **Test and measurement tryout data collected**

The test and measurement tryout data that should be collected includes:

Scores on the criterion tests associated with instructional units, modules, lessons, or lesson segments.

Scores on the criterion post-tests.

Scores on the diagnostic post-tests.

The error rate on practice items or exercises contained in the instructional materials

---

#### **Measuring acceptability of instructional components: Satisfactory**

A component of the instructional material can be considered satisfactory and does not require revision if it achieves the established test and measurement standards for:

The criterion tests and post-tests.

The practice items or exercises included in the instructional materials.

---

**Measuring  
acceptability of  
instructional  
components:  
Unsatisfactory**

A component of the instructional material can be considered unsatisfactory and does require revision if it achieves one of the following two types of results:

Type A: The error rate on practice items or exercises contained in the instructional materials is satisfactory, but performance on the criterion tests and post-tests is below standard.

Type B: The error rate on both the practice items or exercises contained in the instructional materials and on the criterion tests and post-tests is below standard.

In the Type A problem, the error rate is satisfactory on the practice items or exercises contained in the instructional materials, but unsatisfactory on the criterion tests and post-tests.

In the Type B problem, the error rate is below standard on both the practice items or exercises contained in the instructional materials and on the criterion tests and post-tests.

**Type A failures:  
Performance on  
criterion is below  
standard**

Instructional Materials should be revised if either Type A or Type B problems are discovered during the internal review process. The following tables describe some common reasons for Type A or Type B failures:

**Type A Failures**

Failure in Retention	Failure in Transfer
<p>The student correctly performed on test and measurement instruments administered during learning of the instructional materials, but failed to remember what was learned during instruction on the criterion tests.</p>	<p>The student correctly performed on test and measurement instruments administered during learning of the instructional materials, but failed to apply what was learned to a similar situation not previously encountered during instruction on the criterion tests or post-tests.</p>

**Type B failures:  
Performance on  
everything is  
below standard**

**Type B Failures**

<b>Failure in Acquisition</b>
The student failed to learn the material during instruction, and demonstrated that failure by making errors during administration of test and measurement instruments contained in the instructional materials and also during administration of the criterion tests and post-tests.

**Revising  
instructional  
materials:  
Failure in  
retention**

After determining the instructional component where student failures occurred, the types of student errors should be analyzed to determine which corrective measures should be taken to revise the instructional materials. The following table indicates the diagnostic checks and revision treatments that should be considered for failures in student retention of skills or knowledge:

<b>Failure in Retention</b>
<p>Is the criterion behavior actually practiced unassisted? Check to see that the practice items or exercises do not provide excessive help cues for the student. Be sure that the help cues are faded in the instruction to enable students to perform practice items or exercises under criterion conditions. Ensure that the instructional materials include multiple unassisted practice items or exercises associated with specific skills. Check to see that there are multiple practice opportunities for each objective. If practice opportunities are limited, add more practice of each skill or knowledge associated with each objective. Check to see if practice is distributed throughout the course of instruction. Ensure that the learning sequence is long enough, and that adequate reviews are provided. If adequate reviews are not provided, add more practice opportunities and reviews. Include a review of the key teaching points at the end of each lesson to improve retention on the criterion tests and post-tests.</p>

**Revising  
instructional  
materials: Failure  
in transfer**

After determining the instructional component where student failures occurred, the types of student errors should be analyzed to determine which corrective measures should be taken to revise the instructional materials. The following table indicates the diagnostic checks and revision treatments that should be considered for failures in student transfer of skills or knowledge:

<b>Failure in Transfer</b>
<p>Ensure that there are sufficient examples included in the instructional materials.</p> <p>Check to see if practice has been distributed across the full range of conditions stated in the objectives.</p> <p>Check to see if students mistakenly identified non-examples of a class as examples of a class on classification test items (failure in classification of concrete or defined concept properties).</p> <p>Check to see if students mistakenly identified an example of a class as being a non-example of a class on classification test items (failure in classification of concrete or defined concept properties).</p> <p>Check to see if all critical properties that are the basis for classifications and all incidental (non-relevant) properties of a class have been covered in the instructional material. If required, revise examples of critical and incidental properties of concrete and defined concepts, and add instructional materials to cover all relevant properties of the concepts.</p> <p>Check to see if both positive and negative examples of concrete and defined concepts are included in the instructional materials. At least 50% of the examples should be negative examples.</p> <p>Add negative examples of concrete and defined concepts to the instructional materials as required to help the students discriminate between members and non-members of a class.</p> <p>Ensure that the presentation and practice modes of the instructional materials are similar to the actual criterion test item performance.</p>

**Revising instructional materials: Failure in transfer (Continued)**

**Failure in Transfer (Continued)**

Ensure that the mode of presentation and practice in the instructional materials is the same as the student will encounter on the job. For example, if visual perception (discrimination) of cues is required on the job, are similar cues provided in the instructional materials?  
Revise instructional media as necessary to provide the student with job-associated cues to the maximum extent possible. The instruction should simulate job performance skills and knowledge as much as possible.

**Revising instructional materials: Diagnostic checks and revision treatments**

After determining the instructional component where student failures occurred, the types of student errors should be analyzed to determine which corrective measures should be taken to revise the instructional materials. The following table indicates the diagnostic checks and revision treatments that should be considered for failures in student acquisition of skills or knowledge:

**Failure in acquisition**

Causes of failure in acquisition include:  
Lack of practice, or insufficient practice of the criterion behavior to be learned.  
Inadequate feedback about the correctness of student practice items and exercises.  
Diagnostic checks and revision treatments include:  
Check to see if criterion behaviors are actually practiced unassisted.  
Check to see that test and measurement items are not over-cued, and that the cues are faded to provide for unassisted practice of the criterion objectives. Revise test and measurement items if they are over-cued.  
Ensure that there are multiple practice opportunities for each specific skill or knowledge associated with the criterion objectives. Provide more practice if required.  
Check to see if tasks or learning behaviors can be broken down into smaller instructional components to increase student acquisition of the associated skills or knowledge.

---

**Revising instructional materials:  
Diagnostic checks and revision treatments  
(Continued)**

Ensure that sufficient feedback is provided in the instructional materials during student performance of criterion skills and knowledge.

Revise the instructional materials by adding additional cues, exaggeration of differences and similarities, and providing additional rules or principles if student acquisition is a problem.

Check to see if practice or exercises in the instructional materials are appropriate for the type of learning stated in the criterion objectives.

---

**Revision guidelines:  
Learning conditions**

The instructional materials may have failed to educate or train the student to criterion standards because of intrinsic problems with the instructional materials. The following table describes seven learning conditions that cause difficulty for students, and the type of learning affected:

---

Table 3 Learning Conditions That Create Difficulty for Students

Source of Difficulty	Description	Type of learning affected
Complex Instructional Component Parts and Lack of Student Practice or Embedded Testing of Terminal or Enabling Learning Behaviors	The student cannot practice or be tested on the terminal or enabling learning behaviors stated in an objective. The more complex the instructional components are, the more difficult it will be to learn the material associated with an objective, and to master the objective to criterion standards.	Discriminations Memorization Components Concrete Concepts Defined Concepts Rule Learning Verbal Information Cognitive Strategies
Interference From Previous Learning	The student has previously learned a behavioral response to an instructional condition which interferes with the learning of a new response to the condition. The stronger the old response, the harder it will be for the student to learn the new response. The old response will be resistant to change.	Discriminations Memorization Components Concrete Concepts Defined Concepts Rule Learning Verbal Information Cognitive Strategies
Length of Memorization of Chaining Components of Learning Behaviors	The student is required by the objectives to memorize long lists of procedural steps and actions, verbatim rules, etc. The longer the memorization component, the harder it is for the student to meet the criterion objective.	Memorization Components Rule Learning Cognitive Strategies
Number of Cognitive Information Processing Inputs, Processes, and Outputs	The more cognitive information processing inputs, processes, and outputs associated with a learning behavior, there are to be associated, discriminated, or classified, the harder it will be to (1) perceive and encode the inputs, processes, and outputs, (2) make behavioral judgments and decisions in response to the inputs, processes and outputs, or (3) to discriminate the essential similarity or dissimilarity of the inputs, processes and outputs, and (4) to make judgments about the results expected upon completion of a learning behavior.	Discriminations Concrete Concepts Defined Concepts Rule Learning Cognitive Strategies

Table 3 Learning Conditions That Create Difficulty for Students (Continued)

Source of Difficulty	Description	Type of learning affected
Number of Cognitive Information Processing Inputs, Processes, and Outputs	The more cognitive information processing inputs, processes, and outputs associated with a learning behavior, there are to be associated, discriminated, or classified, the harder it will be to (1) perceive and encode the inputs, processes, and outputs, (2) make behavioral judgments and decisions in response to the inputs, processes and outputs, or (3) to discriminate the essential similarity or dissimilarity of the inputs, processes and outputs, and (4) to make judgments about the results expected upon completion of a learning behavior.	Discriminations Concrete Concepts Defined Concepts Rule Learning Cognitive Strategies
Similarity of Cognitive Information Processing Inputs, Processes, and Outputs	The greater the similarity between cognitive information processing inputs, processes, and outputs, associated with a learning behavior, the harder it is for the student to perceive and encode their differences. The student may make an incorrect encoding judgment and make a decision to perform an incorrect behavior as a result.	Discriminations Concrete Concepts Defined Concepts Rule Learning Cognitive Strategies
Dissimilarity of Cognitive Information Processing Inputs, Processes, and Outputs	The greater the apparent dissimilarity between cognitive information processing inputs, processes, and outputs, associated with a learning behavior, the harder it is for the student to perceive and encode their essential similarity. The student may make an incorrect encoding judgment and make a decision to perform an incorrect behavior as a result.	Discriminations Concrete Concepts Defined Concepts Rule Learning (Rule-Using and Problem-Solving) Cognitive Strategies
Large Number of Attributes for Cognitive Information Processing Inputs, Processes, and Outputs	The more attributes that there are to perceive and encode for the cognitive information processing inputs, processes, and outputs, associated with a learning behavior, the harder it is for the student to see their essential similarity. (For example, during test and measurement, it is harder to rate a student on 10 attributes than it is to rate the student on 3 attributes.)	Discriminations Concrete Concepts Defined Concepts Rule Learning Cognitive Strategies

**Revision  
guidelines:  
Learning  
procedural skills**

The following table provides guidelines for overcoming student difficulties in learning procedural skills:

Table 4 Guidelines for Overcoming Student Difficulties in Learning Procedural Skills

Difficulty	Guidelines
Length of Memorization Components of Procedural Segments	<p>Provide several demonstrations of procedural performance or material to be memorized. Point out relevant cues that the student should be aware of.</p> <p>Provide cues that the student can use during practice. For example, a series of photographic, graphic, audio, or video cues can be included in the instructional materials. These media can be digitized for computer-based interactive courseware.</p> <p>Use backward chaining to help the student learn long sets of memorized items, such as procedural steps and actions. To use backward chaining to learn the steps in a procedure, the steps are learned in reverse order, and the student practices the steps in the normal order to the end of the procedure after each step is presented in reverse order by the instruction.</p> <p>Observing the completion of the procedure may help to reinforce student learning of the procedure.</p>
Interference From Previous Learning	<p>Provide frequent opportunities for practice. The amount of cueing should be reduced slowly if interference exists.</p> <p>If the identity of the interference can be determined, make the student aware of it. Find an appropriate means to discourage or extinguish the interference and encourage or reinforce the desired behavior.</p>
Complex Instructional Component Parts and Lack of Student Practice or Embedded Testing of Terminal or Enabling Learning Behaviors	<p>Provide for the student to learn the component part of a psychomotor or intellectual skill before exposing the student to the entire skill or knowledge.</p> <p>Provide instruction on all component parts of a psychomotor or intellectual skill.</p> <p>Provide practice and embedded test items on all component parts of a psychomotor or intellectual skill.</p>

**Revision  
guidelines:  
Learning verbal  
information**

The following table provides guidelines for overcoming student difficulties in learning verbal information (memorization component) skills.

Table 5 Guidelines for Overcoming Student Difficulties in Learning Verbal Information (Memorization Component) Skills

<b>Difficulty</b>	<b>Guidelines</b>
Large Number of Attributes for Cognitive Information Processing Inputs, Processes, and Outputs	Provide a pre-education/training session on similar instructional material and fade the associated cues slowly during the pre-educating/training session. Cluster or chunk the attributes of the cognitive information processing inputs, processes, and outputs of performance into groups that are meaningful to the student.
Length of Memorization Components of Learning Behaviors	Divide the memorization component into meaningful parts and provide practice on the parts. Gradually combine parts until the entire memorization component is learned.
Interference From Previous Learning	Cue potential differences in confusing material by using a series of photographic, graphic, audio, or video cues in the instructional materials. These media can be digitized for computer-based interactive courseware.
Complex Instructional Component Parts and Lack of Student Practice or Embedded Testing of Terminal or Enabling Learning Behaviors	Provide for the student to learn the verbal memorization component part of a psychomotor or intellectual skill before exposing the student to the entire skill or knowledge. Provide instruction on all verbal memorization component parts of a psychomotor or intellectual skill. Provide practice and embedded test items on all verbal memorization component parts of a psychomotor or intellectual skill.

**Revision  
guidelines:  
Learning  
discrimination  
intellectual skills**

The following table provides guidelines for overcoming student difficulties in learning discrimination intellectual skills.

Table 6 Guidelines for Overcoming Student Difficulties in Learning Discrimination Intellectual Skills

Difficulty	Guidelines
Large Number of Attributes for Cognitive Information Processing Inputs, Processes, and Outputs	Provide multiple practice opportunities for each input, process, and output. Then present all inputs, processes, and outputs in operational succession so that students can see the essential differences between the inputs, processes, and outputs.
Similarity of Cognitive Information Processing Inputs, Processes, and Outputs	Exaggerate differences. Ask the student to define the basis for the identity of an input, process, or output.
Large Number of Cognitive Information Processing Inputs, Processes, and Outputs	Reduce the number of inputs, processes and outputs presented to the student early in the instruction, and gradually require the student to cognitively process more and more input, process, and output variables until the student is performing at criterion level.
Interference From Previous Learning	Cue the differences between previous discriminations or concepts, and current discriminations and concepts. Increase practice opportunities.
Complex Instructional Component Parts and Lack of Student Practice or Embedded Testing of Terminal or Enabling Learning Behaviors	Provide for the student to learn the discriminations associated with a component part of a psychomotor or intellectual skill before exposing the student to the entire skill or knowledge. Provide instruction on discriminations associated with all component parts of a psychomotor or intellectual skill. Provide practice and embedded test items on discriminations associated with all component parts of a psychomotor or intellectual skill.

**Revision guidelines:  
Learning defined concept**

The following table provides guidelines for overcoming student difficulties in learning concrete and defined concept (classification) intellectual skills.

Table 7 Guidelines for Overcoming Student Difficulty in Learning Concrete and Defined Concept (Classification) Intellectual Skills

Difficulty	Guidelines
Large Number of Attributes for Cognitive Information Processing Inputs, Processes, and Outputs	Provide practice opportunities across the full range of concept attributes associated with each input, process, and output. Present examples across the full range of concept attributes associated with each input, process, and output in close succession so students can see the essential similarities between concepts associated with the inputs, processes, and outputs.
Dissimilarity of Cognitive Information Processing Inputs, Processes, and Outputs	Exaggerate similarities. Ask the student to provide a definition of each concept.
Large Number of Cognitive Information Processing Inputs, Processes, and Outputs	Reduce the number of concept attributes to be learned, if possible. Focus on the most important concept attributes early in the instruction. Gradually add concept attributes until criterion performance is achieved. Highlight important concept attributes with cues or other attention-getting devices.
Interference From Previous Learning	Provide an overview or advance organizer in the instruction that calls attention to differences between previously learned concepts and the new concepts in the instructional materials. Ask the student to define the concepts.
Complex Instructional Component Parts and Lack of Student Practice or Embedded Testing of Learning Behaviors	Before undertaking the learning of a concept, a student must have learned the associated discriminations and verbal associations. Provide opportunities for practice and testing of associated discriminations and verbal associations before providing instruction on a concept.

**Revision guidelines:  
Learning rules and  
problem-solving**

The following table provides guidelines for overcoming student difficulties in learning rule-learning intellectual skills.

Table 8 Guidelines for Overcoming Student Difficulties in Learning Rule-Learning Intellectual Skills

Difficulty	Guidelines
<p>Large Number of Cognitive Information Processing Inputs, Processes, and Outputs</p>	<p>If the rule-using or problem solving skill requires learning of a large number of inputs, processes, and outputs, associated with a procedure, provide instruction on the more critical elements before combining all of the elements.</p> <p>If the rule-using or problem-solving skill requires the combining of a large number of procedures including the tasks, subtasks, steps and actions associated with each procedure, provide instruction on one procedure at a time before combining all of the procedures.</p>
<p>Interference From Previous Learning</p>	<p>Provide an overview or advance organizer in the instruction that calls attention to differences between previously learned rules or problem solving procedures and the new rules or problem-solving procedures in the instructional materials. Ask the student to define the rules or problem-solving steps, actions, and results expected.</p>
<p>Complex Instructional Component Parts and Lack of Student Practice or Embedded Testing of Terminal or Enabling Learning Behaviors</p>	<p>Provide for the student to learn the component parts of a rule-using or problem-solving procedure before exposing the student to the entire rule or problem-solving procedure.</p> <p>Provide instruction on all component parts of a rule-using or problem-solving procedure.</p> <p>Provide practice and embedded test items on all rule-using or problem-solving procedure components.</p>

**Revision guidelines:  
Forming desired attitudes**

The following table provides guidelines for overcoming student difficulties in forming desired attitudes.

Table 9 Guidelines for Overcoming Student Difficulties in Forming Desired Attitudes

<b>Difficulty</b>	<b>Guidelines</b>
Interference From Previous Learning	Describe undesirable attitudes, and why they are undesirable. Follow with a description of the desirable attitude and why it is desired, with emphasis on the benefit to the student. Provide rewards for the student for manifestations of desired attitudes.

---

## Section F

### Conducting a Validation of the Instructional System

---

#### Introduction

The last two steps in the *internal review (formative evaluation)* process are to:

Subject the instructional system to *small group-tryouts*.  
Subject the instructional system to an *operational tryout*. *The operational tryout is the final step in the validation process and is the summative evaluation of the instructional materials.*

---

#### Small-group tryout guidelines

*Small-group tryouts* should be conducted on a sample of the student population that is representative of the entire student population that will enter the instructional system. The *small-group tryout* should be the next to last step in the internal review (formative evaluation) process, and should be conducted prior to conducting the operational tryout (summative evaluation).

The *individual and single-group tryouts* and the revisions that were made to the instructional materials as a result of the initial steps of the formative evaluation process should have been sufficient to produce an acceptable instructional system for the small group tryouts.

The *individual and single-group tryouts* should have been conducted using a sample of the student population that came from the upper 25% in aptitude and background.

The *small-group tryouts* of the instructional system are conducted using a sample of the student population that is representative of the entire student population that will enter the instructional system.

---

#### Selecting a representative sample and conducting small-group tryouts

The following guidelines should be followed to select a representative student sample for small-group tryouts:

Select students who represent a sample of the total student population.

The sample should include an even distribution of low, average, and high-aptitude students.

---

**Selecting a representative sample and conducting small-group tryouts (Continued)**

Select 6-10 students, try out the instructional system on them, and make revisions. Then try out the instructional system on 6-10 additional students. Continue the small-group sampling until a total of 20-30 typical prospective students have been exposed to the instructional system.

**Small group tryouts: Types of data**

The following table describes the types of data that should be gathered during small-group tryouts.

Error rate data	Time data
Record the accuracy of student responses to test items and test and measurement instruments in the same manner that was used for the individual and single-group tryouts.	Record the time needed by each student to complete each instructional unit, module, lesson, and lesson segment. Record the time needed by each student to complete each test item and test and measurement instrument.

**Small group tryouts: Measuring time data**

Time is a factor in the small-group tryout phase of the validation process. The following considerations should be observed when measuring time data:

It is not sufficient that a student is able to learn the instructional material.

The student must also complete instructional system components in a reasonable time. Use both time and error metrics.

In self-paced components of instructional systems administered outside of a classroom or learning center, there may not be a limit placed on the length of time a student can take to master the objectives. For example, interactive computer-based courseware used as a *desk-top reference*. This is a desired function for this type of instruction.

In self-paced components of *Criterion-Referenced instructional systems* administered in a classroom or learning center, there must be a limit placed on the length of time a student can take to master the objectives.

---

**Small group tryouts: Measuring time data (Continued)**

In group-paced components of *Criterion-Referenced instructional systems* administered in a classroom, learning center, on-the-job environment, etc., there must be a limit placed on the length of time a student can take to master the objectives.

Efforts to establish time limits for an instructional component should be based upon the instructional requirements stated in the objectives and the capabilities of the majority of the students in the group.

---

**Small group tryouts: gathering time data**

The following considerations should be observed for gathering time data:

Calculate the *median* completion time needed by each student to complete each instructional unit, module, lesson, and lesson segment.

Calculate the *median* completion time needed by each student to complete each test item and test and measurement instrument.

The median time is important, because further revision of the instructional system may be required to reduce or increase completion times for instructional components.

It is not practical to pace instruction to meet the needs of the slowest or fastest student.

The estimates of instructional time required that is gathered during the small-group tryouts should be based upon the observed median times of student performance.

---

**Small group tryouts: Gathering error rate data**

The following considerations should be observed for gathering error rate data:

In the small-group tryout, the validation effort should include a wider range of student aptitudes than the individual and single-group tryouts.

Identify any error pattern that occurs in any lesson segment in the course of instruction.

Take steps to strengthen any lesson segment where error patterns occur.

Use the same method for isolating lesson segments that need revision as was used in the individual and single-group tryouts.

---

---

**Small group tryouts: Gathering error rate data (Continued)**

If a lesson segment requires a significant revision, repeat the small-group tryout of the lesson segment after the revision has been made.

Give the revised instruction to small groups of 6-10 students, until data have been gathered on 20-30 more students.

Continue the small-group tryouts until the students can perform to the level specified in the objectives and the test and measurement instruments.

---

**Conducting the operational tryout and implementation (summative evaluation)**

The operational tryout should be conducted by personnel who will be the administrators and instructors in the operational system. The following considerations apply to the operational tryout:

Test complete instructional sequences on approximately 30 students who represent the student population.

The length of the sequence will vary. A unit, module, block, lesson or lesson segment may be subjected to summative evaluation during operational tryout.

If possible, the entire course of instruction should be evaluated.

Time and resource availability will dictate the scope of the instructional area that is subjected to operational tryout.

---

**Reasons for conducting an operational tryout**

Reasons for conducting an operational tryout include the following:

Instructional materials, including test and measurement instruments must be evaluated as an integral part of the total instructional system.

Individual, single group, and small group tryouts are used to validate instruction in an isolated environment. The operational tryout validates the instructional system in an operational environment.

Analysis of data from the larger student sample used during operational tryout will provide a solid base for final revision and refinement of the instructional system.

---

---

**Reasons for conducting an operational tryout (continued)**

Data gathered from the operational tryout will provide feedback about the adequacy of the task or learning analysis, the terminal and enabling objectives, the test items, the test and measurement instruments, and the instructional content and presentation.

If students fail to meet the objectives during operational tryout, each step in the instructional system development process should be analyzed to determine the source of the problem.

An operational tryout encompasses the evaluation of administrative, equipment, training device, computer, facility and any other instructional resources associated with the instructional system.

---

**Reasons for conducting summative evaluations**

The requirement to continually conduct summative evaluations and to make revisions to an instructional system is based on the following factors:

In addition to problems with instructional materials, there may be problems with instructor qualifications, equipment, training devices, computers, maintenance, scheduling, and variations in student attitudes and aptitude.

Changes in job performance requirements or job restructuring will probably impose changes on criterion objectives and changes in instructional materials, instructor qualifications, equipment, training devices, computers, maintenance, and scheduling.

The instructional system must be continually revised to meet changing needs of the student population, operational requirements, administrative problems, and management problems.

Improperly developed or invalid revisions can destroy an instructional system by reducing its ability to enable student mastery of valid objectives.

Restrain the impulse to implement a quick fix to an instructional system problem until its impact on the total instructional system and operational job requirements have been analyzed.

---

## Bibliography

- Bills, C.G., and Butterbrodt, V.L. (1992). *Total Training Systems Design Function: A Total Quality Management Application*. Wright-Patterson AFB, Ohio.
- Briggs, L.J., and Wager, W.W. (1981). *Handbook of Procedures for Design of Instruction* (2nd Ed.). Glenview, Illinois: Harper Collins Publishers
- Carlisle, K.E. (1986). *Analyzing Jobs and Tasks*. Englewood Cliffs, New Jersey: Educational Technology Publications.
- Davies, I.K. (1976). *Objectives in Curriculum Design*. London: Mc Graw Hill.
- Dick, W., and Carey, L. (1990). *The Systematic Design of Instruction* (3rd Ed.). Glenview, Illinois: Harper Collins Publishers.
- Gagné, R.M. (1985). *The Conditions of Learning* (4th Ed.). New York: Holt, Rinehart and Winston.
- Gagné, R.M., Briggs, L.J., and Wager, W.W. (1992). *Principles of Instruction* (4th Ed.). New York: Harcourt Brace Jovanovitch College Publishers.
- Gagné, R.M., and Merrill, M.D. (1990). *Integrative Goals for Instructional Design*. Englewood Cliffs, New Jersey: Educational Technology Publications. 38(1), 1-8.
- Goldstein, I.L. (1986). *Training In Organizations: Needs Assessment, Development, and Evaluation* (2nd Ed.). Pacific Grove, California. Brooks/Cole Publishing Company.
- Hageman, D.C. (1988). *Cognitive Engineering of Training Systems for Simulators*. National Aerospace and Electronics Conference. Dayton, Ohio.
- Hageman, D.C. (1985). *Effective Training Systems for High-Technology Equipment Operation*. National Security Industrial Association Fifth Annual Conference on Personnel and Training System Effectiveness. San Antonio, Texas.
- Keller, J.M. (1987). The Systematic Process of Motivational Design." *Performance and Instruction*, 26(9), 1-8.
- Kibler, R.J. (1981). *Objectives for Instruction*. Boston: Allyn and Bacon.
- Knirk, F.G., and Gustafson, K.L. (1986). *Instructional Technology: A Systematic Approach to Education*. New York: Holt, Rinehart, and Winston.
- Leshin, C.B., Pollock, J., and Riegeluth, C.M. (1992). *Instructional Design Strategies and Tactics*. Englewood Cliffs, New Jersey: Educational Technology Publications.
- Mager, R.F. (1962). *Preparing Objectives for Instruction* (2nd Ed.). Belmont, California: Fearon Publishers.
- Merrill, M.D., Tennyson, R.D., and Posey, L. (1992). *Instructional Design Strategies and Tactics*. Englewood Cliffs, New Jersey: Educational Technology Publications.
- Merrill, M.D., Lee, Z., and Jones, M.K. (1990). *Second Generation Instructional Design*. Englewood Cliffs, New Jersey: Educational Technology Publications.
- O'Neil, H.F., Jr., and Baker, E.L. (1991). Issues in Intelligent Computer-Assisted Instruction: Evaluation and Measurement." In T. Gutkin and S. Wise (Eds.), *The Computer and the Decision Making Process*. Hillsdale, New Jersey: Erlbaum Lawrence Associates.
- Reigeluth, C.M. (1983). Instructional design; what is it and why is it? In C.M. Reigeluth (Ed.), *Instructional Design Theories and Models? An Overview of Their Current Status*. Hillsdale, New Jersey: Erlbaum Associates.
- Rossett, A. (1987). *Training Needs Assessment*. Englewood Cliffs, New Jersey: Educational Technology Publications.
- Spears, W.D. (1983). *Processes of Skill Performance: A Foundation for the Design and Use of Training Equipment*. (NAVTRAEQ-VIPCEN 78-C-0113-4). Orlando, Florida: Naval Training Equipment Center.

**Bibliography (Continued)**

Tennyson, R.D., and Michaels, M. (1991). *Foundations of Educational Technology; Past, Present and Future*. Englewood Cliffs, New Jersey: Educational Technology Publications.

Wolfe, P. et.al. (1991). *Job Task Analysis: Guide to Good Practice*. Englewood Cliffs, New Jersey: Educational Technology Publications.

## Chapter 5

# GUIDELINES FOR OPERATIONAL EVALUATION OF INSTRUCTIONAL RESOURCES

---

### Purpose of this chapter

The information in this chapter is to be used in conjunction with the information contained in AFM 36-2234, Instructional System Development, and in AFH 36-2235, Information for Designers of Instructional Systems, Volumes 1-11.

The purposes of this chapter are to:

Provide guidelines for the internal evaluation of instructional system resources, including the definition and purpose, the possible causes for problems, data collection and analysis, and reporting the findings.

Provide guidelines for the external evaluation of instructional system resources, including the definition and purpose, the possible causes for problems, data collection and analysis, and reporting the findings.

Provide guidelines for field visits, including the definition and purpose, the possible causes for problems, data collection and analysis, and reporting the findings.

Provide guidelines for job performance evaluation, including the definition and purpose, the possible causes for problems, data collection and analysis, and reporting the findings.

### Where to read about it

This chapter contains five sections.

Section	Title	Page
A	Introduction	132
B	Internal Evaluation	136
C	External (Field) Evaluation	144
D	Questionnaires	146
E	Field Visits	153
F	Job Performance Evaluation	157

---

## Section A Introduction

---

### Introduction

After the instructional materials (including test and measurement instruments and instructional resources) have been validated by individual, single-group, and small group tryouts, and have been subjected to an operational tryout (*summative evaluation*), the instructional system is ready for implementation.

Once the instructional system has been implemented and starts producing graduates, it is time to begin conducting *operational evaluations*.

Operational evaluation is a continuous process, and assesses how well course graduates are meeting the established *job performance requirements*.

---

### Overview of operational evaluation

Evaluation of the instructional system components and resources is a continuous process, starting with *summative evaluation* (validation of instructional materials including lesson plans, instructor and student guides, test and measurement instruments, and instructional resources, such as equipment, training devices, interactive courseware, audiovisual media, hardcopy media, facilities, manpower, costs, instructors, and administrators).

Summative evaluation is conducted through individual, single-group, and small-group tryouts, and lastly by an operational tryout.

---

### Definition of operational evaluation

The last stage of the evaluation process is *operational evaluation*. Operational evaluation is the continuous process of gathering and analyzing *internal and external feedback data* about the instructional system to ensure that the system continues to effectively and cost-efficiently produce graduates who meet established *job performance requirements*. Operational evaluation is a quality improvement operation.

---

---

**Purpose of operational evaluation**

There are two main purposes of operational evaluation:

- Ensure that graduates continue to meet established job performance requirements.
  - Continually improve instructional system quality.
- 

**What to look for during operational evaluation**

Look for both strengths and weaknesses in the instructional system during operational evaluation. Operational evaluation should focus on:

- How well the graduates are meeting job performance requirements.
  - Whether instruction is being provided that is not needed.
  - Whether instruction that is needed is not being provided.
  - How well each instructional system component is contributing to overall instructional system quality. Instructional components include lesson plans, instructor and student guides, workbooks, reference materials, test and measurement instruments, equipment, training devices, interactive courseware, audiovisual media, Plans of Instruction (POI), Course Training Standards (CTS), syllabi, facilities, manpower, costs, instructors, and administrators.
  - Ways to improve the graduate's performance as well as the instructional system.
- 

**Operational evaluation activities: Internal evaluation**

There are two operational evaluation activities:

**Internal Evaluation**

- The internal evaluation process gathers internal feedback and management data from the instructional system during operation in the actual education/training environment.
- The data are gathered to assess the effectiveness and quality of the instructional system in its operational environment.

**External (Field) Evaluation**

- The external evaluation process is also called field evaluation because the process gathers data from the field to assess graduate's on-the-job performance in an operational job environment.
  - Most external evaluation data are gathered by evaluators from the organization providing the instruction, or are provided directly from the graduates or their supervisors in the field.
-

---

**Operational  
evaluation  
activities: External  
evaluation  
(Continued)**

In some cases, external evaluation data are gathered and provided to the organization conducting the instruction by both Air Force and Non-Air Force consultants, advisory bodies, accrediting agencies, and other professional education/training certification groups. (i.e., the Federal Aviation Agency).

---

## **Section B**

### **Internal Evaluation**

---

**Overview of internal evaluation**

Internal evaluation activities begin with implementation of the instructional system following internal review (validation) and operational tryout (summative evaluation) of the instructional system.

Internal evaluation activities continue throughout the life cycle of the instructional system.

Sometimes, internal evaluation is called a “course review” by some organizations.

Internal evaluations analyze the instructional system during operation in the actual education/training environment to determine system effectiveness and quality.

---

**Definition of internal evaluation**

Internal evaluation is the acquisition and analysis of internal instructional system feedback and management data such as:

Test and measurement data.

Student critique data.

Instructor comment data.

Data correlation from within the instructional system.

---

**Purpose of internal evaluation**

The purpose of internal evaluation is to improve the quality and effectiveness of the instructional system.

---

**Possible causes for problems during internal evaluation**

Even though an instructional system has been validated before implementation, students may still have difficulty with the instruction during day-to-day instructional system operation. Some of the causes of student problems with the instruction include:

Instructors do not follow the Plan of Instruction (POI), the Course Training Standards (CTS), or the Course Syllabus. The course of instruction that was developed is different in some respects from the course that has been actually implemented.

---

**Possible causes for problems during internal evaluation (Continued)**

Instructional resources that are required to support, operate, and maintain the course of instruction differ in some respects from the resources that have been actually implemented. The actual instructional resources used in the course of instruction are inadequate for the students to master specific terminal or enabling objectives. Instructional materials are not correlated with the test and measurement instruments, with the terminal or enabling objectives, or with the instructional content identified in the task and learning analyses. Students do not match established course prerequisites.

**Data collection**

The following table describes the purposes of several methods of collecting internal evaluation data.

<b>Data Collection Methods</b>	<b>Purpose of Data Collection</b>
Review Course Control Documents	To determine if there are any discrepancies between the planned course and the course that was actually implemented.
Review Resources	<p>To ensure that facilities (instructional and support) are available.</p> <p>To ensure that equipment and training devices (instructional, support, and test and measurement), and supplies are available.</p> <p>To ensure that human resources (instructional developers, instructors, students, and courseware maintenance personnel) are available.</p> <p>To ensure that there is adequate time allotted for the instruction (adequate course length and sufficient time to maintain the course).</p> <p>To ensure that adequate funds are available to support, operate, and maintain the course.</p>
Visit Instructional Facilities	<p>To evaluate the quality of implemented instruction. Ensure that the visit is long enough to observe samples of representative instruction for the entire course.</p> <p>To check hardcopy instructional materials such as instructor and student guides, workbooks, and reference materials for quality and availability.</p> <p>To check equipment, training devices, instructional media, and training aids for condition, operation, and appropriateness.</p>

**Data collection  
(Continued)**

<p>Evaluate Instructor Performance</p>	<p>To check if the instructors follow the Plan of Instruction, Course Training Standards and Syllabus. To check if the instructors use instructional media properly, respond to student needs, and are qualified to teach. To check instructor evaluation forms to see if noted weaknesses have been corrected.</p>
<p>Monitor Test and Measurement Standards</p>	<p>To check the test and measurement program for compromise. If a test or measurement instrument has been compromised, it cannot provide useful feedback on student performance. To monitor the test and measurement program to ensure quality of the test and measurement items and student performance. To evaluate instruction in terms of student performance. Test and measurement instruments are the performance measures that determine student achievement of course objectives.</p>

**Conducting an internal evaluation**

Collect sufficient internal evaluation data for the analysis. Insufficient data may skew the analysis results, and possibly lead to incorrect decisions being made based on the results of the internal evaluation.

Job aids can be used to gather internal evaluation data. The following is an example of a job aid for internal evaluation.

Check	Internal Evaluation Data Source
	Does the POI/CTS/Syllabus reflect the operational course?
	Is the POI/CTS/Syllabus current and accurate?
	Does the POI/CTS/Syllabus provide adequate guidance?
	Do the lesson plans, workbooks, instructor guides, student guides, and reference materials and the POI/CTS/Syllabus agree?
	Do the lesson plans, workbooks, instructor guides, student guides, and reference materials reflect what is being taught in the course?

**Conducting an  
internal evaluation  
(Continued)**

Check	Internal Evaluation Data Source
	Are the lesson plans, workbooks, instructor guides, student guides, and reference materials current and accurate?
	Do the instructional media, and instructional materials support the POI/CTS/Syllabus?
	Do instructional facilities meet system requirements?
	Does instructional equipment support the POI/CTS/Syllabus and meet system requirements?
	Do training devices support the POI/CTS/Syllabus and meet system requirements?
	Do support facilities meet system requirements?
	Does training equipment meet system requirements, and is it adequately maintained?
	Do the training devices meet system requirements, and are they adequately maintained?
	Are instructors teaching according to the lesson plans?
	Have instructors received training on the purpose and execution of the course of instruction?
	Have the instructors been adequately trained?
	Do the test and measurement instruments adequately measure the terminal and enabling objectives?
	Is test and measurement data thoroughly analyzed?
	Can improvements be made in the course?

**Example of job aid:  
Classroom  
instruction**

Job aids can be used to gather internal evaluation data. The following is an example of a job aid for obtaining student reaction to classroom instruction:

### Student Reaction Job Aid

COURSE OF

INSTRUCTION \_\_\_\_\_ DATE \_\_\_\_\_

INSTRUCTOR \_\_\_\_\_ STUDENT \_\_\_\_\_

One way instruction is improved is by sampling student reaction to the instruction. Please answer the following questions:

1. Prior to this instruction, my experience in this area was <input type="checkbox"/> extensive <input type="checkbox"/> moderate <input type="checkbox"/> little or none	7. Audiovisual aids were <input type="checkbox"/> of great value <input type="checkbox"/> valuable <input type="checkbox"/> of little or no value <input type="checkbox"/> not used, but could be of help <input type="checkbox"/> not used, and not needed
2. Did your knowledge of the instruction increase as a result of the instruction? <input type="checkbox"/> yes <input type="checkbox"/> no	8. Answers to student questions were <input type="checkbox"/> meaningful <input type="checkbox"/> somewhat helpful <input type="checkbox"/> not helpful <input type="checkbox"/> not applicable (no questions)
3. If your knowledge increased as a result of the instruction, to what extent did it increase? <input type="checkbox"/> not applicable (no increase) <input type="checkbox"/> slightly <input type="checkbox"/> moderately <input type="checkbox"/> extremely	9. Should the subject matter be changed? <input type="checkbox"/> yes (please explain below) <input type="checkbox"/> no
4. Based on my experience, the level of the instruction was <input type="checkbox"/> too advanced <input type="checkbox"/> about right <input type="checkbox"/> too elementary	10. Should the method of instruction be changed? <input type="checkbox"/> yes (please explain below) <input type="checkbox"/> no
5. The organization of the instruction was <input type="checkbox"/> very helpful <input type="checkbox"/> helpful <input type="checkbox"/> not very helpful	11. Overall, the instruction was <input type="checkbox"/> outstanding <input type="checkbox"/> good <input type="checkbox"/> fair <input type="checkbox"/> poor
6. The lesson outline (student guide) was <input type="checkbox"/> very helpful <input type="checkbox"/> helpful <input type="checkbox"/> not very helpful	12. Test and measurement instruments were <input type="checkbox"/> outstanding <input type="checkbox"/> good <input type="checkbox"/> fair <input type="checkbox"/> poor

**Example of job aid:  
Interactive  
courseware**

Job aids can be used to gather internal evaluation data. The following two pages are an example of a job aid for obtaining student reaction to interactive courseware (ICW):

COURSE OF

INSTRUCTION \_\_\_\_\_ DATE \_\_\_\_\_

INSTRUCTOR \_\_\_\_\_ STUDENT \_\_\_\_\_

One way instruction is improved is by sampling student reaction to the instruction. Please answer the following questions:

	Item	SD	D	A	SA
1	All function keys/ screen prompts worked correctly.				
2	The Help instructions were clear and easy to follow.				
3	All branching mechanisms worked correctly (Back, Forward, Continue, Help, Media, Glossary, Menu, Quit)				
4	Graphics had clarity and supported the learning objectives .				
5	Videos had clarity and supported the learning objectives.				
6	Text was clear and supported the learning objectives.				
7	Grammar was correct and appropriate for the material being presented.				
8	Technical jargon was appropriate for the material being presented.				
9	All acronyms and abbreviations were properly identified.				
10	The program was easy to use.				
11	Lesson material was sequenced appropriately.				
12	When desired, access to any lesson segment or lesson screen worked correctly.				
13	All words were spelled correctly.				

Comments by screen number are mandatory for all items rated D or SD.  
Enter your comments on the next page.

Key:  
SD - Strongly Disagree  
D - Disagree  
A - Agree  
SA Strongly Agree

---



---

**Internal evaluation data analysis:  
Methods of analyzing**

Some methods of analyzing the internal evaluation data are:

Compare the instructional lesson or lesson segment terminal and enabling objective standards with the standards in the Plan of Instruction (POI)/Course Training Standard (CTS)/Syllabus to determine if the requirements of the POI/CTS/Syllabus are being met.

Compare the POI/CTS/Syllabus with the operational course to determine if the planned and operational courses are the same.

Review the POI/CTS/Syllabus and the lesson plans, instructor guides, student guides, and other instructional materials to see if they are current, adequate, and in agreement.

Compare stated instructional resource requirements with actual resources to determine if adequate resources are available to support, operate, and maintain the instructional system.

Review instructor records to ensure instructors are qualified to teach the course of instruction.

Review test and measurement data to determine if students are meeting the terminal and enabling objectives.

Analyze test and measurement instruments to determine if test and measurement items are valid and reliable.

---

**Revising the instructional system following internal evaluation**

After internal evaluation data are collected and analyzed, the next stage is to correct deficiencies in the instructional system. If revisions can be made to correct identified problems, they should be made in a timely manner to achieve the greatest benefit from the revisions.

Revisions resulting from the internal operational evaluation analysis may require returning to the analysis or design phase of the instructional systems development process, depending on the scope of the revision.

Changing a test item, or adding instructional time to a unit of instruction would not require returning to the analysis or design phase of the instructional systems development process, but adding a new piece of equipment to the course would require analysis.

---

## Section C

### External (Field) Evaluation

---

**Introduction**

The purpose of *external evaluation* is to determine how well graduates of an instructional system are meeting job performance requirements. External evaluation relies on input from the operational job environment (field) in order to determine how well the graduates are performing.

---

**Definition**

External (field) evaluation is the process of gathering and analyzing data from outside the instructional environment in order to determine how well recent graduates are meeting job performance requirements.

---

**Purpose**

The purpose of external evaluation is to determine if recent graduates of the course:

- Can meet job performance requirements.
  - Need all of the instruction they received.
  - Need any instruction they did not receive.
- 

**Possible causes of problems**

Possible problems that may be identified during external evaluation include:

- Criterion test(s) do not measure graduates ability to meet job performance requirements.
  - Terminal or enabling objectives do not reflect job performance requirements.
  - Job performance requirements were incorrectly identified during task and learning analyses.
  - Job performance requirements changed after task and learning analyses.
-

---

**Collecting data**

Several methods of collecting external evaluation data are listed in the following table, and addressed in subsequent sections of this chapter:

<b>Methods of Internal Evaluation</b>	<b>Page</b>
Questionnaires	146
Field Visits	153
Job Performance Evaluation	157
Other Sources of Evaluation Input	159

---

## **Section D**

### **Questionnaires**

---

#### **Introduction**

Questionnaires are effective, cost-efficient evaluation tools. This section will address the following topics:

- Advantages and disadvantages of questionnaires.
- Types of questionnaires.
- How to prepare and distribute questionnaires.
- Analysis of data gathered using questionnaires.

---

#### **Purpose of questionnaires**

The purpose of using questionnaires is to:

- Determine the ability of recent graduates to perform specific tasks on which they received instruction to the standards stated in the terminal and enabling objectives.
- Identify the specific nature of any deficiency in the instructional system.
- Determine what tasks are actually being performed by graduates.
- Identify what components of the instructional system are not required to educate/train personnel for performance of actual job tasks.

---

#### **Advantages of questionnaires**

Advantages of questionnaires include:

- They are comparatively inexpensive to administer.
- They can be used to collect large samples of data from course graduates and their supervisors.
- They yield data that can be easily tabulated and reported.
- Respondents usually give their opinions freely.

---

#### **Disadvantages of questionnaires**

Disadvantages of questionnaires include:

- They may not be the most reliable form of evaluation. The validity of the data depends on how the questionnaires are prepared and distributed.
- Communication is one way (to the respondent). The respondent may not understand some of the questions.

---

---

**Disadvantages of questionnaires (Continued)**

Questionnaires collect only opinion data. Therefore, questionnaires may not be as reliable as other forms of collecting external data.  
 Developing effective and reliable questionnaires may be costly and require extensive experience.  
 Low return rates for questionnaires will affect reliability and validity.  
 Inappropriate responses will affect accuracy.

---

**Types of questionnaires**

Two types of questionnaires can be used to collect external evaluation data:

One type of questionnaire is for the graduates' immediate supervisor. The response may be delegated to the graduates' instructor instead of the supervisor.  
 The other type of questionnaire is for the graduates. This type of questionnaire is designed to find out what the graduates think about the instruction they received.

---

**Preparing questionnaires**

Well constructed questionnaires that are properly administered are extremely important to the external evaluation process. The following table identifies the five basic stages of questionnaire development.

<b>Stage</b>	<b>Activity</b>
Stage 1	Define the purpose of the questionnaire. Focus only on relevant information.
Stage 2	Determine the specific information to be collected. Specify exactly what is needed in a list of topics and sub-topics.
Stage 3	Develop questionnaires that ask for specific information such as: What conditions and equipment are required to perform the job. Exact actions required to accomplish a job task. Standards of job task performance. Conditions for job task performance. Expected results of job task performance.

---

**Preparing questionnaires  
(Continued)**

Stage	Activity
Stage 4	Consider motivational factors when developing questionnaires. You want the respondents to answer fully and conscientiously. Questionnaires should motivate if they: Explain the purpose of the questionnaire. Tell the respondents how they can benefit from answering the questionnaire. Contain clear and concise instructions. Have an uncluttered format and are easy to answer. For example, using boxes for check marks. Have questions arranged in a logical order. Contain specific questions.
Stage 5	Test the questionnaire on sample respondents. Ask them to: Evaluate the cover letter. Check instructions and questions for clarity. Explain how they feel about answering the questions. Revise the questionnaire, if necessary, before distribution.

**Guidelines for developing effective questions**

Guidelines for developing effective questions include the following:

Use *closed-end* questions when you want the respondent to choose answers from a small number of possibilities. Closed-end questions makes tabulation of the responses easy, but may not provide the range of answers required.

Use *open-end* questions when you don't know all the possible answers. The respondent will probably suggest possibilities.

Word questions to the respondent's level of understanding.

Use vocabulary and concepts that are easy for the respondent to understand.

Limit each question to one aspect of a topic.

Decide on the logical order of the questions. For example, list questions by task performance order, from general to specific.

---

**Guidelines for developing effective questions (Continued)**

Avoid questions that make the desired answer obvious.  
 Avoid questions that show bias, state opinions or contain exceptions.  
 Word questions so that they will not threaten the respondents.  
 Supplemental questions designed to obtain additional information may be used. For example, questions that ask how much time a graduate spends on individual tasks, or what equipment or materials the graduate uses may be used to obtain additional information.

---

**Guidelines for constructing questionnaires**

When constructing a questionnaire, consider these guidelines:

Provide short, concise, and specific directions for completing the questionnaire. The directions should be printed in heavy bold type, if possible.  
 Provide space for the respondent's name, title, organization, and location.  
 Number the questionnaires for administrative control.

Whenever possible, allow the respondent to use the same type of marking to answer all questions. For example, one of the best methods is to allow check marks for responses.

Arrange "yes" and "no" answers vertically rather than horizontally.

**Correct**

Yes \_\_\_\_\_  
 No \_\_\_\_\_

**Incorrect**

Yes \_\_\_\_ No \_\_\_\_

Number each page of the questionnaire.  
 The questionnaire should be easy to read and mark.  
 The questionnaire should be printed.  
 Print on both sides of the pages to conserve materials, if possible.  
 Send a self-addressed return envelope with the questionnaire.  
 Fold the questionnaire in such a manner that the respondent can refold it the same way to place it in the return envelope after completion.

---

---

**Guidelines for preparing cover letters**

Each questionnaire should have a cover letter. Ensure that the cover letter:

- Explains the purpose of the questionnaire and its importance for improving instruction.
  - Includes a statement that ensures the respondent the information will be treated confidentially.
  - Includes a statement that the evaluation is being conducted in accordance with applicable Air Force directives.
  - Provides information on how to return the questionnaire.
  - Indicates the approximate time required to complete the questionnaire.
  - Shows the date the questionnaire was mailed and the recommended return date.
  - Uses appropriate letterhead stationery signed by a responsible authority.
- 

**Distribution of questionnaires:  
Administer to small sample**

Before distributing the questionnaire, administer it to a small number of select individuals to:

- Provide valuable feedback on the quality of the questionnaire.
  - Preclude acquiring misinformation resulting from the administration of a faulty questionnaire.
  - Allow correction of problems in the questionnaire before distribution.
- 

**Guidelines for distribution of questionnaires**

Distribution of the questionnaires is a critical aspect of external evaluation. The following are guidelines for distributing the questionnaire:

- Plan the distribution to ensure that the data collected are both valid and reliable.
  - Decide to whom you are sending the questionnaire — to recent graduates, supervisors, instructors or a combination of personnel.
  - Select a representative sample to ensure valid results.
  - Graduates may perform different tasks, or their job requirements may vary depending upon their major command, geographic location, or organizational level. Questionnaires should be distributed to graduates at each major command and organizational level.
-

**Determining the number of questionnaires**

Determine how many questionnaires are needed to be mailed out. Base this decision on:

Expected response return rate.

Level of confidence of the sample. This is a statistical consideration that specifies how large a sample is required to state with a percent of confidence (e.g., 90% confident) that the sample is statistically representative of the larger population. The following table shows the number of graduates that must be sampled for a given confidence level.

**Graduate Sampling Chart**

Course Graduates During Sampling Period	Sample Size 95% Confidence	Sample Size 90% Confidence	Sample Size 80% Confidence
10	10	10	9
20	19	19	18
40	36	35	32
60	52	49	44
80	67	62	54
100	80	73	62

**Example of Use of Table**

Annual *course production* is 100 graduates.

The *confidence level* selected is 95%. (It is recommended that the 95% confidence level be selected, since this is the level commonly used in business decisions.)

*Estimated return rate* for the questionnaires is 85%.

$$\frac{85\%}{100\%} = \frac{80}{X}$$

$$X = \frac{80 \times 1.00}{.85} = 94 = \text{number of questionnaires to mail}$$

---

**Determining when to distribute questionnaires**

Decide when to distribute the questionnaires. Timing is critical.

Usually, questionnaires should be sent to the graduates within three to six months after graduation.

If the questionnaire is sent too late, it may be impossible to tell whether the graduate learned the skill or knowledge in the course of instruction or on the job.

If the questionnaire is sent too early, the graduate may not have had time to perform many of the tasks or apply much of the knowledge that was contained in the course of instruction.

---

**Questionnaire data analysis**

When a sufficient number of completed questionnaires have been returned, the external evaluation data are analyzed. The data are:

*Compiled* by major command, geographic location, or organization level.

*Collated* by major command, geographic location, or organization level.

*Analyzed* by major command, geographic location, or organization level.

During analysis of the external evaluation data, pay particular attention to:

Notes made by respondents on the questionnaires.

Answers to any supplemental questions that were included in the questionnaire.

Carefully check the data to ensure that data with the following obvious errors are used with caution or not used at all during data analysis:

Halo Effect: indiscriminate rating of all items positively.

Central Tendency: indiscriminate grouping of rated items in the center of a scale.

Examine the responses to ensure, insofar as possible, that the information accurately reflects the opinion of the graduates and their supervisors.

---

## **Section E**

### **Field Visits**

---

#### **Introduction**

Field visits are a very effective method of conducting external evaluations. Field visits consist of interviews of graduates from a course of instruction, and observation of their performance on the job.

Field visits are normally conducted by an evaluator, often assisted by an instructional developer or instructor.

Field visit evaluators should include individuals who are specialists in the job area being evaluated, and who are familiar with the jobs the graduates are performing.

Field visit constraints include limited funds, scheduling of evaluators, and the number and variety of graduates to be interviewed and observed on the job.

---

#### **Purpose of field visits**

The purpose of a field visit is to get first-hand information on the graduates' assignment, utilization, and proficiency on the job. Another purpose is to validate information gained from previous evaluation activities such as formative and summative evaluation.

---

#### **Advantages of field visits**

Advantages of field visits are:

Guidance and information about the evaluation is given directly to the graduates, their instructors, or their supervisors. Information is gathered first-hand by the evaluators. Any questions or assumptions can be clarified on the spot. Field visits can help validate previous questionnaire data. External evaluations, especially field visits, help establish rapport between the instructional organization and the operational end user of the instructional system graduate. Additional information can be obtained from the graduates by observing nonverbal messages and by asking leading or probing questions.

---

---

**Disadvantages of field visits**

Disadvantages of field visits are:

They are time-consuming. For example, travel may be required to several bases to evaluate an appropriate sample of graduates. Also, interviews and observations of performance under job conditions require a great deal of time if they are accomplished correctly.

Usually, due to time, scheduling, and money constraints, the sample of graduates is limited. The evaluator may only go to a few bases, and the number of interviews and observations of performance under job conditions may also be limited.

The cost is high. Field visits require evaluators to spend funds for travel to the various bases.

Interview and observation data gathered by the evaluator can be subjective and biased.

Graduates may feel that they are being scrutinized.

Evaluators have not always received training on the conduct of evaluations, and therefore may not be skilled at interviewing and observing.

---

**Data collection during field visits**

The two methods of collecting data during field visits are:

Interviews  
Observations

Evaluators should interview recent graduates and their instructors or supervisors. Observations will be almost useless unless the observer is familiar with the tasks to be performed and the standards for task performance.

---

**Preparing for the field visit**

Visits to the field to collect evaluation data should be adequately planned to ensure that useful data are gathered. To prepare for field visits, develop a list of questions that will elicit honest, pertinent answers, and that will keep the interview focused. Then, accomplish the following:

Determine the bases to be visited.  
Establish the schedule for the visit.  
Select the individuals to be interviewed and observed.

---

---

**Conducting the field visit tasks**

The following are some of the tasks to be performed during the field visit:

**Conduct Interviews**

Inform graduates, instructors, and supervisors of the purpose of the visit. Tell them that their answers will furnish valuable information for improving the instruction.

Interview the recent graduates and their instructors or supervisors. Instructors or supervisors should know how well the graduate has performed on the job.

Guide the interviews using the prepared list of questions. As the interview progresses, additional questions may need to be asked, or questions may need to be revised or deleted.

Take accurate and complete notes, especially on information that is freely given.

Determine the graduate's level of proficiency for performance of job tasks and associated intellectual skills.

**Document Performance**

Find out how the graduates are progressing in their On-the-Job Training (OJT) program.

Determine how the intellectual skills, motor skills, and attitudes learned during instruction are being used.

Observe the graduates as they are performing tasks. The evaluator should have prior job and task knowledge.

Take careful notes on graduate performance. After each task has been completed, ask questions to clarify actions taken by the graduate during task performance.

Have supervisors or instructors rate the graduates' performance.

---

**Field visit data analysis**

Data collected from field visit interviews and observations are analyzed in the same manner as questionnaire data. The data should be:

*Compiled* by major command, geographic location, or organization level.

*Collated* by major command, geographic location, or organization level.

*Analyzed* by major command, geographic location, or organization level.

---

---

**Reporting the field  
visit findings**

The results of the external evaluation gathered by field visits and questionnaires should be combined and reported in the Training Quality Report (TQR).

The data gathered during field visits are not normally used or reported independently from the questionnaire data.

The results of the analysis of the questionnaire data and the field visit data are compared in order to validate the findings.

---

## Section F

### Job Performance Evaluation

---

**Introduction**

Job performance evaluations are a type of external evaluation that is conducted jointly by the instructional organization and a *using command*.

Job performance evaluations are conducted in the operational job environment.

Job performance evaluations are based on a using command's Job Performance Requirements (JPR).

---

**Purpose of job performance evaluations**

The purpose of job performance evaluations is to determine how well recent graduates of a course of instruction are meeting the *using command's job performance requirements (JPR)*.

---

**Advantages of job performance evaluations**

Advantages of job performance evaluations include the following:

Evaluations are conducted on the job by the supervisor.

Evaluations are very thorough.

The supervisor submits reports on a weekly basis. This procedure ensures an accurate assessment over time of the graduate's performance.

Job performance evaluation data can be used to validate other forms of external (field) evaluations such as the interview and the questionnaire.

---

**Disadvantages of job performance evaluations**

Disadvantages of job performance evaluations include the following:

It usually takes 8 to 10 weeks to conduct the evaluation.

Supervisor time is required to make the weekly reports.

The evaluator makes at least two trips to each base to collect data.

The graduate sample is limited.

Job performance evaluations normally focus on a single command.

---

---

**Job performance evaluation data collection**

Job performance evaluation data are collected via field reports submitted by a supervisor to an evaluation agency. The weekly job performance evaluation reports summarize the progress made by a graduate during the previous week.

---

**Preparing for the job performance evaluation**

As with any evaluation method, make adequate plans before starting. Planning tasks for job performance evaluations include the following:

Select recent graduates and their supervisors to participate in the job performance evaluation.

Meet with the supervisor and the graduates to explain job performance evaluations, and get the supervisor's commitment to support the evaluation.

Determine the tasks to be evaluated based on the major command training standard. *The criterion for task performance is the major command training standard.*

Establish evaluation milestones for the 8 to 10 week evaluation period.

---

**Conducting job performance evaluations**

Once the participants have been selected and briefed on the job performance evaluation process, the evaluation can begin. The job performance evaluation consists of the following activities:

The supervisor evaluates and records the graduates' performance on each task performed.

The supervisor reports the following on a weekly basis:

Tasks performed.

Frequency of task performance.

Time required to perform the tasks.

Equipment used to perform the tasks.

---

**Job performance: Evaluation data analysis and reporting**

When the evaluator receives the job performance reports from the supervisor, the reports are analyzed to determine how well the graduates are performing the tasks they were taught in the course. Evaluators should watch for reports that indicate the graduate:

Can not perform a task that was taught in the course.

---

**Job performance:  
Evaluation data  
analysis and  
reporting  
(Continued)**

Requires excessive help to perform a task that was taught in the course.

If a graduate cannot perform a task or requires excessive help to perform a task that was taught in the course of instruction, the data analysis should focus on determining why the graduate is not able to meet the major command job performance requirements.

Job performance evaluation is normally conducted in conjunction with other forms of external (field) evaluations (questionnaires/interviews), the results of job performance evaluations are included in the Training Quality Report (TQR).

**External evaluation  
data: Inspector  
General (IG)  
reports**

Some other external evaluation data sources that can be used to evaluate graduates job performance include:

*Inspector General (IG) Reports*

The Air Force and the major commands periodically inspect instructional and operational activities to determine their effectiveness.

The inspections conducted by these inspection teams may discover problems related to external evaluation.

Use IG Report data to determine if graduates are meeting their job performance requirements.

Take appropriate action to correct deficiencies identified in IG Reports.

One example of an IG Report is the Functional Management Inspection.

**External evaluation  
data:  
Standardization  
evaluation reports**

Standardization/Evaluation teams periodically inspect instructional and operational activities to determine their effectiveness and compliance with standards.

Analyze findings indicating a problem, and take appropriate action to correct the deficiencies.

---

**External evaluation  
data: Air Force  
training quality  
reports**

*AF Form 1284, Training Quality Report (TQR)*

The TQR contains external evaluation data from interviews, questionnaires, and job performance evaluations.

The TQR reports strengths and weaknesses of instruction that the graduates received.

The instructional activity should respond to any deficiencies identified in the TQR. Use problems identified in the TQR to validate external evaluation data.

---

## **Bibliography**

- Bills, C.G., and Butterbrodt, V.L. (1992). *Total Training Systems Design Function: A Total Quality Management Application*. Wright-Patterson AFB, Ohio.
- Briggs, L.J., and Wager, W.W. (1981). *Handbook of Procedures for Design of Instruction* (2nd Ed.). Glenview, Illinois: Harper Collins Publishers
- Carlisle, K.E. (1986). *Analyzing Jobs and Tasks*. Englewood Cliffs, New Jersey: Educational Technology Publications.
- Davies, I.K. (1976). *Objectives in Curriculum Design*. London: Mc Graw Hill.
- Dick, W., and Carey, L. (1990). *The Systematic Design of Instruction* (3rd Ed.). Glenview, Illinois: Harper Collins Publishers.
- Gagné, R.M. (1985). *The Conditions of Learning* (4th Ed.). New York: Holt, Rinehart and Winston.
- Gagné, R.M., Briggs, L.J., and Wager, W.W. (1992). *Principles of Instruction* (4th Ed.). New York: Harcourt Brace Jovanovitch College Publishers.
- Gagné, R.M., and Merrill, M.D. (1990). *Integrative Goals for Instructional Design*. Englewood Cliffs, New Jersey: Educational Technology Publications. 38(1), 1-8.
- Goldstein, I.L. (1986). *Training In Organizations: Needs Assessment, Development, and Evaluation* (2nd Ed.). Pacific Grove, California. Brooks/Cole Publishing Company.
- Hageman, D.C. (1988). *Cognitive Engineering of Training Systems for Simulators*. National Aerospace and Electronics Conference. Dayton, Ohio.
- Hageman, D.C. (1985). *Effective Training Systems for High-Technology Equipment Operation*. National Security Industrial Association Fifth Annual Conference on Personnel and Training System Effectiveness. San Antonio, Texas.
- Keller, J.M. (1987). The Systematic Process of Motivational Design. *Performance and Instruction*, 26(9), 1-8.
- Kibler, R.J. (1981). *Objectives for Instruction*. Boston: Allyn and Bacon.
- Knirk, F.G., and Gustafson, K.L. (1986). *Instructional Technology: A Systematic Approach to Education*. New York: Holt, Rinehart, and Winston.
- Leshin, C.B., Pollock, J., and Riegeluth, C.M. (1992). *Instructional Design Strategies and Tactics*. Englewood Cliffs, New Jersey: Educational Technology Publications.
- Mager, R.F. (1962). *Preparing Objectives for Instruction* (2nd Ed.). Belmont, California: Fearon Publishers.
- Merrill, M.D., Tennyson, R.D., and Posey, L. (1992). *Instructional Design Strategies and Tactics*. Englewood Cliffs, New Jersey: Educational Technology Publications.

## Bibliography (Continued)

- Merrill, M.D., Lee, Z., and Jones, M.K. (1990). *Second Generation Instructional Design*. Englewood Cliffs, New Jersey: Educational Technology Publications.
- O'Neil, H.F., Jr., and Baker, E.L. (1991). Issues in Intelligent Computer-Assisted Instruction: Evaluation and Measurement. In T. Gutkin and S. Wise (Eds.), *The Computer and the Decision Making Process*. Hillsdale, New Jersey: Erlbaum Lawrence Associates.
- Reigeluth, C.M. (1983). Instructional Design; What is it and Why is it? In C.M. Reigeluth (Ed.), *Instructional Design Theories and Models? An Overview of Their Current Status*. Hillsdale, New Jersey: Erlbaum Associates.
- Rossett, A. (1987). *Training Needs Assessment*. Englewood Cliffs, New Jersey: Educational Technology Publications.
- Spears, W.D. (1983). *Processes of Skill Performance: A Foundation for the Design and Use of Training Equipment*. (NAVTRAEQ-VIPCEN 78-C-0113-4). Orlando, Florida: Naval Training Equipment Center.
- Tennyson, R.D., and Michaels, M. (1991). *Foundations of Educational Technology; Past, Present and Future*. Englewood Cliffs, New Jersey: Educational Technology Publications.
- Wolfe, P. et.al. (1991). *Job Task Analysis: Guide to Good Practice*. Englewood Cliffs, New Jersey: Educational Technology Publications.

## Chapter 6

### ADVANCED MEASUREMENT TOPICS

---

#### Purpose of this chapter

The information in this chapter is to be used as advanced supplemental information that expands on the information contained in the first five chapters of this handbook.

The purposes of this chapter are to:

Provide guidelines for standard-setting for Norm-Referenced and Criterion-Referenced tests, including descriptions of standard-setting procedures.

Provide guidelines for Criterion-Referenced test analysis, including indices of test reliability and item discrimination.

Provide guidelines for test validation, and lessons learned for test item development.

#### Where to read about it

This chapter contains four sections.

Section	Title	Page
A	Guidelines for Standard-Setting for Norm-referenced and Criterion-referenced Tests	164
B	Criterion-referenced Test Analysis	183
C	Guidelines for Test Validation	195
D	Lessons Learned for Test Item Development	202

#### Primary reference

The advanced measurement topics discussed in this chapter are based on the following documents produced by Dr. Thomas R. Renckly of the Air University:

Renckly, T.R. (April 1993) Practitioner's Guide to Standard-setting. Air University, Montgomery, AL.

Renckly, T.R. (October 1990) Criterion-Referenced Test Analysis: Another Look at a Compromised Process. Paper presented at the 14th Annual Inter-Service Correspondence Exchange Conference, Pensacola, FL.

---

---

**Primary reference  
(Continued)**

Renckly, T.R. (April 1990) A Guide to Criterion-Referenced Test Analysis. Air University, Montgomery, AL.

Renckly, T.R. (October 1989) Test Validation: Current Practice—New Perspectives. Paper presented at the 1989 Inter-Service Exchange Conference, Williamsburg, VA.

---

## **Section A**

### **Guidelines for Standard-Setting for Norm-Referenced and Criterion-referenced Tests**

---

#### **Deciding on the process for standard-setting**

The first step in setting passing standards for Norm-Referenced and Criterion-Referenced tests is to determine the standard-setting process that you want to use. To develop rational passing standards you need a commitment to a standard-setting process and a knowledge of the steps of the process. Each standard-setting process is:

- Designed to serve a particular purpose
- Based on a particular set of assumptions
- Requires that certain kinds of data be collected, and
- Necessitates a unique interpretation of the resulting passing score.

---

#### **Steps for standard-setting**

The steps in standard-setting are:

- Deciding on the purpose for the standard-setting method.
- Selecting the method for standard-setting, and the steps involved in using the method.
- Selecting judges.
- Defining the metric against which judgments will be made.
- Training judges.

---

#### **The first step in standard-setting: Deciding on the purpose for the standard-setting method**

The method selected for setting passing standards will depend upon the data that are available. The ultimate purpose for a standard-setting method must be known to be able to select the most appropriate method.

If you know the types of data available, you can select the standard-setting method based on the types of available data. If the types of data are not known or available, the standard-setting method can be selected based on the purpose of the standard you are trying to set.

---

---

**Standard setting:  
Nedelsky, Angoff,  
and Ebel methods**

*Nedelsky, Angoff, and Ebel Methods*

These methods are *based on expert judgments about test questions*.

These methods are easy to implement and do not require observation or evaluation of actual student performance. They can be used either before or after administration of a test.

These methods share the same purpose--to determine a *standard that will distinguish borderline performance from inadequate or failing performance on a test*.

None of the methods considers the performance of fully qualified students. The only interest is how borderline students perform. (*Distinguish borderline performance from inadequate or failing performance on a test.*)

The validity of any of these three methods depends to a large degree on the definition of *borderline performance as determined by judges*.

---

**Standard setting:  
Borderline group  
method**

*Borderline Group Method*

This method is based on *experts' judgments about individual test-takers (students)*.

This method is also concerned with the *performance of borderline students*.

This method makes use of *actual student performance, not hypothetical judgments* of student performance.

The purpose of this method is also to *distinguish borderline performance from inadequate or failing performance on a test*.

*The validity of this method is high, because the resulting standard is based on actual student performance, and therefore could be replicated in actual testing situations.*

A similar *Instructed-Uninstructed Groups Method* does not rely on expert judgment about group differences, but relies on statistical validity and classification probabilities.

---

**Standard setting:  
Contrasting groups  
method**

*Contrasting Groups Method*

The purpose of this method is to establish a *standard* which distinguishes between qualified and unqualified individuals. The *degree* to which an individual is qualified or unqualified is not of interest.

Only the *division line between qualified and unqualified is important*.

A much finer distinction must be made to determine a student's abilities as borderline compared to simply distinguishing between qualified or unqualified.

Judging whether a student is qualified or unqualified *does not require a knowledge of the precise level of performance, or a minimal level of performance*.

**Standard setting:  
Reference group  
method**

*Reference Group Method*

The purpose of this method is to define a *passing standard in relation to actual performance of a reference group*.

This is the *most realistic method and should always be used when establishing passing standards*, if possible.

The data requirements of this method are the most rigorous.

The required data are expensive to gather in terms of time and level of effort.

This is a very viable method and produces *results that are highly valid and reliable compared to the other methods*.

**The second step:  
Selecting steps for  
standard-setting  
methods using  
Nedelsky's,  
Angoff's, and  
Ebel's methods**

*Steps for Accomplishing Nedelsky's, Angoff's, and Ebel's Methods (Methods Based on Making Judgments About Test Questions)*

Select judges.

Define "borderline" knowledge and skills.

Train judges in the method to be used.

Collect, combine and compute judgments.

Choose a passing score.

**The second step:  
Steps for standard  
setting using  
borderline,  
contrasting,  
instructed, and  
uninstructed  
groups methods**

*Steps for Accomplishing the Borderline-Group, Contrasting  
Groups, and Instructed-Uninstructed Group Methods  
(Methods Based on Judgments About Individual Test Takers)*

Select judges (not done in the Instructed-Uninstructed Groups Method).

Define “adequate”, “inadequate”, and “borderline” levels of knowledge and skills.

Identify actual “borderline”, or “qualified” and “unqualified”, or “instructed” and “uninstructed” test takers.

Obtain test scores for all members of each identified group. Choose a passing score. In the Contrasting Groups Method and the Instructed-Uninstructed Groups Method, the passing score is determined by the point of overlap between each group’s test score distribution.

**The second step:  
Steps for standard  
setting using  
reference group  
method**

*Reference Group Method  
(Method Based on Judgments About a Group of Test-Takers)*

Select judges.

Identify a reference group.

Define “adequate” and “inadequate” levels of knowledge and skills.

Collect and compute judgments.

Choose a passing score.

**Commonality of  
procedural steps**

All of the methods, except the Instructed-Uninstructed Groups Method, have two procedural steps in common:

Selection of judges (experts) who will provide information that is useful in setting the ultimate passing standard and defining the decision metric that will be used to differentiate between test takers on one side of the passing cutoff score or the other. All the methods choose a passing score as their last step, but the process is accomplished differently depending on the method selected.

**The third step in  
standard-setting:  
Selecting expert  
judges**

Who Will Be Chosen? (Who Are the Experts?)

Characteristics of experts in a domain of knowledge or skill:

Experts tend to excel in their own domains.  
 Experts are able to perceive meaningful patterns in their domain of expertise.  
 Experts are able to perform skills rapidly.  
 Experts represent problems at a deep principled level.  
 Experts spend time analyzing problems qualitatively before they begin to take action on them.  
 Experts tend to have strong self-monitoring skills; know when they need more information; and know when they are proceeding incorrectly.  
 Experts tend to be more accurate at judging problem difficulty than non-experts.

**Expert  
characteristics for  
judging test  
questions**

Where Can Experts Be Found?

*For Methods Based on Judgments About Test Questions, Look for Experts With These Characteristics*

Knowledgeable of the content domain (subject matter) covered by the test questions.  
 Ability to determine the relative difficulty level of each test item.  
 Able to reasonably speculate about a student's ability to answer each test item correctly.  
 Experience as test developers or classroom teachers.

**Expert  
characteristics for  
judging individual  
test-takers**

*For Methods Based on Judgments About Individual Test-Takers, Look for Experts With These Characteristics*

Individuals who can judge the skill performance of test takers.  
 Ability to differentiate between qualified and unqualified students in terms of their skill performance and their adequate and inadequate levels of knowledge in a particular domain.  
 Individuals who have observed the students performing the skills under consideration.  
 Experience as teachers as well as job supervisors.

---

**Expert characteristics for judging group test-takers**

*For Methods Based on Judgments About A Group of Test-Takers, Look for Experts With These Characteristics*

Familiar with the skills and abilities of the reference group.  
Be able to determine the level of knowledge and skills possessed by the members of the group that relate to the test under consideration.

Be familiar with what the test measures.

Experience as test developers, teachers, and job supervisors.

Or, experience as job supervisors exclusively, and also with the contents of the test(s) being considered in the standard-setting process.

---

**Procedures for selecting judges**

*How Should Experts Be Selected?*

Administering a competency examination to the judges may raise additional standard-setting issues.

It may be more reasonable to identify the judge's knowledge after they have provided their judgments using van der Linden's (1982) *intra-judge consistency metric*, or Sato's (1975) *caution index* post hoc analysis techniques.

Sato's caution index is more widely applicable, and easier to use and interpret.

The index is based on the notion of an ideal response pattern in an individual's test answers.

It is unreasonable to expect all experts to judge every question on a test identically.

It is not unreasonable to expect all experts' judgments to be fairly consistent. This is called an *ideal response pattern*.

If some judges' ratings appear to systematically deviate from the mean rating ascribed by all the judges, those judges may be identified as relative novices.

---

---

**How many experts?***How Many Experts Should Be Chosen?*

The upper limit on the number of experts is the largest number the facilitator can comfortably manage.

The lower limit on the number of judges should be enough to allow synergy between the experts. A group of less than six people makes it difficult to generate and maintain a discussion.

It is important to get a variety of perspectives when establishing a test standard. If the number of judges is too small, the standard-setting process can be influenced by the opinions of one or two judges whose ratings may be extremely high or low in relation to the rest of the group.

The larger the group, the less effect a single judge's ratings will have on the overall average.

Consider using a group size of 10-20 judges. Break the group into smaller groups, and average the smaller-group ratings, if required for manageability.

The greater the criticality of the standard to be set, the greater the need to have a representative sampling of judges, and the greater the need that judges' rating errors and variations be kept below the chance level of statistical significance. For critical standard-setting, it is not unusual to use 70, 80, or more judges.

---

**The fourth step in standard-setting: judgment metrics***The Specific Metric Used for Making Judgments Varies According to the Standard-Setting Method*

In Nedelsky's, Angoff's, and Ebel's methods (based on judgments about test questions), the metric is the borderline (minimal) level of skill competence as measured on the test. In the Borderline Group and Contrasting Groups Methods (which are based on judgments about individual test takers), the metric is adequate versus inadequate (qualified versus unqualified) levels of skill or knowledge.

The Reference Group Method (which is based on judgments about a group of test takers) uses the same metrics as the Borderline Group and Contrasting Groups Method.

---

---

**The fourth step in standard-setting: validity of a standard**

The validity of a standard is directly dependent on the accuracy and universal acceptance of the rating metric among all the judges.

If some judges are confused or in disagreement about the definition of “borderline” or “adequate” or “qualified”, this will lead to a greater variation in the ratings, and consequently, a greater degree of error.

This error will be random, and will not be statistically adjusted. There is no way to determine to what degree this systematic error is affecting the ratings.

The validity and reliability of the resulting standard is lowered.

---

**The fourth step in standard setting: Reducing judges' errors**

To reduce systematic errors in judge's ratings, it is vitally important to ensure that they all operate with a consistent definition of the rating metric to be used.

Spend the time to define the metric to be used and the method used for passing standard-setting.

The definition of the metric can be a face value definition. The minimal level can be all that is necessary to graduate from a course or to be promoted, for example. The problem is that minimal performance or minimal competence levels can be established for experienced practitioners, or beginning practitioners, or average practitioners, etc.

A deeper-level definition of the metric overcomes the face value confusion by explicitly considering the metric as part of the definition.

For example, to define competence on the job, a description of the typical entry levels and the minimal acceptable levels for the beginner should be provided to the judges.

For example, definitions of the level of competency required for entry into a job, or to satisfactorily complete the test under consideration should be provided to the judges.

For example, descriptions of the level of knowledge and skills of the typical test-taker and the minimal level for passing the test should be provided to the judges.

---

---

**The fourth step in  
standard-setting:  
Development of  
test standards**

*Suggested Steps for Initial Development of Test Standards*

Judges should begin by developing a generic list of knowledge and skills for an occupation without concentrating on whether they are required for entry level or required for a higher level of expertise.

For job competency tests, ask experts, for example, "What does the typical practitioner do and need to know for graduation or for entry into a profession?"

For tests within a education or training system ask experts, for example, "What does the typical student need to know to pass this examination or to be promoted into the next grade?" Divide the test under consideration into major content or skill areas, and perform the questioning procedure for each part of the test.

After generic knowledge and skill requirements are determined for the typical practitioner or student, the next step is to refine the list to focus more on the entry level into the profession, or minimal competence needed in the next grade, etc.

For example, ask experts "A minimally competent person will at least know the following", and "What types of mistakes are forgivable errors for a minimally competent person?"

The resulting definition of the rating metric should be pared down to a concise, yet complete statement.

The metric may or may not address specific skills but should address the knowledge level that is required to meet the standard.

---

**The fifth step in  
standard-setting:  
Training judges to  
rate**

The purpose of training judges in the standard-setting method to be used is to eliminate misunderstandings that can lead to *inconsistent application of a standard-setting method or rating metric* across many test items or many test takers.

*Training Judges to Rate Test Items*

Provide practice in the Standard-Setting Method using test items that are not on the operational form of the test, but are parallel in form and substance. By analyzing the results of these ratings, the following idiosyncrasies in judges' ratings can be identified:

---

---

**The fifth step in standard-setting: Training judges to rate (Continued)**

Tendency to rate above or below the mean of other judges.  
Rating items inconsistently with respect to item difficulty (e.g., rating a difficult item as easily passable by borderline students).

Conduct periodic retraining by interrupting the standard-setting exercise when inconsistencies or errors are detected in judges' ratings.

Address any idiosyncrasies in the particular Standard-Setting Method used. For example, inform judges of negatively worded test items, multiple true-false items, or items involving computations. These types of test items can pose problems for judges.

---

**Stability of standard setting**

Standard-setting should be stable over time.  
Stability can be measured by using the same or parallel test item at several points throughout the test session.  
A judge's rating should be consistent each time the item is rated.

---

**Consistency of ratings**

Ratings should be consistent with the relative difficulty of items.  
Some of the Standard-Setting Methods use item difficulty as an important data element.  
A judge's ratings should be consistent with item difficulty statistics. If not, a borderline test taker could be judged as being able to pass a high percentage of difficult items or a low percentage of easy items.  
Sato's Caution Index, discussed earlier, can be used to identify the degree to which ratings are consistent with item difficulties.

---

**Ratings reflecting expectations**

Ratings should reflect realistic expectations.  
Some Standard-Setting Methods do not make use of actual student performance data, but use hypothetical estimates about how judges think a minimally competent student might perform on a particular test.

---

---

**Ratings reflecting expectations (Continued)**

These hypothetical estimates may or may not reflect realistic expectations about the population of students who will ultimately take the test.  
Rating items inconsistently with respect to item difficulty (e.g., rating a difficult item as easily passable by borderline students).

Conduct periodic retraining by interrupting the standard-setting exercise when inconsistencies or errors are detected in judges' ratings.

Address and idiosyncrasies in the particular Standard-Setting Method used. For example, inform judges of negatively worded test items, multiple true-false items, or items involving computations. These types of test items can pose problems for judges.

Training is usually accomplished using group discussions and problem analysis.

Training may be computer-based. The courseware would present test items for a judge to rate, and would continue providing practice until the judging criteria for stable standard-setting ratings, consistency with the relative difficulty of test items, and reflection of realistic expectations are met.

---

**Concerns about judges***Concerns About Eliminating Judges Who Don't Meet the Criteria*

The criteria used to measure judges are, like the standard-setting methods themselves, essentially subjective.

Therefore, judges who score items beyond the "norm" are not necessarily wrong.

Judges should have been selected due to their domain expertise and experiential insights and therefore can provide important insights.

If the pool of judges is small, or a large number of judges are excluded, the external validity of the ratings produced is compromised. The results obtained from the judges must be as stable and replicable as possible. It is easier to defend a standard if there is evidence that an independent team of judges would arrive at a similar standard under the same conditions.

---

---

**Using Nedelsky's procedure for standard-setting**

Nedelsky's procedure is used *only with multiple-choice tests* since the method *requires judgment to be made about each possible incorrect answer*. The judge's task is to examine each test question and identify the *incorrect answers that a borderline student would be able to recognize as wrong*. The steps of the process are as follows:

Select judges, and one or more persons to serve as facilitators.

Convene judges and facilitators. Develop a comprehensive definition of what constitutes borderline performance. Put the definition and examples in writing.

Train the judges in the steps of the procedure.

Have judges make a preliminary set of judgments individually for all questions, marking the *incorrect answers a borderline student would be able to eliminate*. Discuss each question, starting with the first. Determine the number of judges who did and did not mark an incorrect answer a borderline student would be able to eliminate.

Obtain explanations for test item answers that do not have unanimous agreement. Emphasize the ratings are for *borderline students*.

Make sure judges have marked all wrong answers they believe a borderline student would be able to eliminate.

Combine judgments and compute results. Choose the passing score.

Correct the passing score for guessing, if desired.

---

**Nedelsky's procedure without and with correction for guessing**

*Without correction for guessing:*

For the first question, eliminate all choices the borderline student would (in the judge's opinion) be able to recognize as correct.

Add the number of choices not eliminated for this question and divide this sum into the number 1. This is the *expected score* for this question.

Continue in the same manner for all test questions. Add the expected scores for all test items. This is the *expected total test score* as determined by a single judge.

Average all judge's expected test scores by adding all scores and dividing by the number of judges. This is the *mean expected test score*.

---

---

**Nedelsky's  
procedure without  
and with correction  
for guessing  
(Continued)**

*With Correction For Guessing*

Add across the entire test all the answers that were not eliminated, and subtract this sum from the total number of test questions. This is the *expected number of wrong answers*. *Divide this value by the number of wrong choices per question*. If test items have different numbers of wrong answers, group those test items with the same number of alternatives together, and consider each group separately from the others.

---

**Using the Angoff  
procedure for  
standard-setting**

The first three steps of this procedure are identical to Nedelsky's. One advantage of Angoff's procedure is that it can handle more than just multiple-choice tests. Judges consider each question as a whole rather than each individual alternative.

*Judges' estimate the probability that a borderline student would be able to answer the question correctly without regard to individual alternatives.* The estimated probability value must be in the range of 0.0 to 1.0.

The easier the question, the higher the probability will be of getting it right. Be sure that the judges estimate for any question is not lower than the probability of guessing the correct answer. This lower (guess-rate) probability limit can be easily calculated by dividing the number of alternatives in the question into 1.

Make and collect judgments. Discuss each question and ask the judges for their *probability estimates* for each question. Probability values should be nearly the same (the highest probability being within 10 to 15 percent (0.1 to 0.15) of the lowest).

Combine the judgments and compute the results. *Choose the passing score*. The passing score can be computed with or without correction for guessing.

---

---

**Angoff procedure  
without correction  
for guessing***Without Correcting for Guessing*

*Average the probability estimates for each question by adding all judge's estimates for a question and dividing by the number of judges.*

*Add all the average probability estimates. The sum is the total expected score for a borderline student (the minimum test passing score).*

---

**Angoff procedure  
with correction for  
guessing***With Correction for Guessing*

If the test items are multiple-choice, true-false, or matching, correction for guessing can be calculated.

Subtract the expected total test score from the total number of test questions. This is the actual score.

Compute the number of penalty points to assess by dividing the actual score by the number of alternatives per question to yield the penalty points.

Subtract the penalty points from the expected total test score to yield the minimum passing score corrected for guessing.

---

**Using the Ebel  
procedure for  
standard-setting**

In Ebel's procedure, judges first classify questions into groups, then estimate the probability of a borderline student answering the questions in each group correctly.

The first three steps are similar to the previous two approaches. In addition, the judges must develop a difficulty-relevance grid. This grid has four relevance columns (Essential, Important, Acceptable, Questionable), and three difficulty rows (Easy, Medium, Hard) for each test item group. Questions are assigned to a cell in the grid.

Make and collect judgments. The judges estimate the percentage of questions a borderline student could answer for each of the cells.

After the judges make preliminary estimates, compare the estimates to ensure that the highest estimate is within 10 to 15 percent of the lowest estimate.

After all the cells are discussed, average the judges' percentage estimates for each cell. Change the percentages to decimals by dividing by 100.

---

---

**Using the Ebel procedure for standard-setting (Continued)**

Combine judgments and compute results--choose the passing score. Multiply each category's percentage (decimal) by the number of questions in that category to yield the expected score for that category.

Add the expected scores for each category to obtain the expected total (minimum passing) test score.

As with the previous standard-setting procedure, the minimum passing test score can be corrected or not corrected for guessing.

---

**Using the borderline-group standard-setting procedure**

Using this procedure, judges identify actual students (test takers) as borderline in knowledge and skills. It is only necessary to judge those students, who, in the judges' opinion, meet the definition of a borderline student. The larger the number of actual test takers identified, the more reliable the resulting passing score will be.

Select judges.

After selecting a suitable number of experts to serve as judges and one or more persons to serve as facilitators, convene all the judges and develop a comprehensive definition of what constitutes a "borderline" performer.

Train judges in the procedure.

Identify actual borderline students.

Make and collect the judgments. Place the test scores on a sheet of paper from lowest to highest and select the score that corresponds to the middle of the distribution (the median test score). This is the minimum passing score.

The judge's scores for the borderline group should cluster together towards the median. If they do not, the borderline group may have several students who do not belong in it. The judges may have classified some of the students incorrectly.

The judges may need to be retrained in what constitutes borderline performance and retrained in the procedure.

---

**Using the contrasting groups procedure**

This procedure is based on the premise that a group of students can be divided into two contrasting groups (qualified and unqualified) according to judgments of their knowledge and skills.

---

---

**Using the  
contrasting groups  
procedure  
(Continued)**

Essentially, this procedure compares students' actual test scores with a judgment of their previous performance on tests of skills and knowledge.

**Phase 1: Define performance levels**

Select judges.

Define adequate (qualified) and inadequate (unqualified) levels of knowledge and skills.

Have judges rate the skill performance of students. Choose as large a group of students as possible. Make sure they are representative of the students who will typically be tested.

After judges have categorized all students as qualified or unqualified, create score intervals that are five or ten points apart.

Collapse the intervals (enlarge the interval size) until all intervals have at least two students in either the qualified or unqualified groups (or both).

Divide the students at each score interval into the qualified and unqualified groups. List the number in a table for ease in computing the passing score.

**Phase 2: Compute the passing score**

Compute the percentage of students at each score interval who are qualified, by dividing the number qualified by the total number of individuals at each score interval.

There will be inversions where a percentage value from a particular interval will be lower than the percentage value of the intervals below it. Use statistical smoothing techniques to eliminate, or significantly decrease, the number of these inversions in preparation for choosing the passing score.

The passing score that is chosen should be based on a consideration of how serious it would be to falsely qualify an otherwise unqualified student or falsely judge a qualified student as unqualified.

If your organization's policy is that these two errors are equally serious, then the passing score of the test should be set at the point where the smoothed percent qualified is exactly 50%.

If, on the other hand, your organization believes it is more serious to pass an unqualified student than it is to fail a qualified one ( a reasonable assumption in safety-dependent job skills and knowledge, for example), then the passing score should be adjusted higher on the percent-qualified scale to perhaps 70% or higher.

---

---

**Using the reference group procedure**

The purpose of this procedure is to develop a passing standard that is validated against an independent group of actual qualified performers. This procedure has the highest degree of validity. However, it is also the most demanding in terms of the data required.

**Phase 1: Identify reference group**

Identify the reference group. The reference group usually consists of the previous year's graduates from the school or course under consideration. It is a relatively straightforward matter to compare their actual postgraduate performance against their previous test scores.

If you use a previous year's group as the reference group, the passing score represents an absolute standard since a student's passing score (qualified or unqualified) is not based on the scores of other students in the current group. The passing score will have been determined before the current group ever took the test.

If you use a current year's group as the reference group, the passing score represents a relative standard since a student's passing score (qualified or unqualified) will be based on the scores of other students in the current group.

**Phase 2: Define performance levels**

Define qualified and unqualified levels of knowledge and skill performance as measured in the test being considered.

Select the judges.

Make and collect judgments. If the judges are present, ask the judges to estimate the percentage of people in the reference group who have an adequate level of the knowledge and skills measured by the test. To do this, provide the judges with a list of the knowledge and skills the test measures for them to refer to as they make their judgments.

If judges are not present, the materials can be provided to the judges in a survey instrument. Survey judges relatively soon after the graduates have arrived on the job, or at the next school. Usually, waiting more than about six months will tend to invalidate a judge's perceptions because the graduate's level of knowledge and skill will grow due to other factors on the job.

---

**Using the reference group procedure (Continued)**

**Phase 3: Determine passing score**

One common way of choosing a passing score is to choose a score that would have passed a particular number (or percentage) of students in the reference group. You can use that score as the passing score.

Choose the passing score. Average the judges' estimates of the number of qualified individuals by adding the estimates and dividing by the number of judges. This will yield the percentage of qualified persons in the reference group. Multiply the total number of students in the reference group by the percentage of qualified persons determined by the judges. Use this numerical value to order the test scores from the highest to the lowest score.

Starting from the highest score, count down until the numerical value previously determined is reached. The score of that student will be the minimum passing score for the test.

**Additional information**

The following bibliography provides additional information on standard-setting.

- Berk, R. A. (Winter 1976) *Determination of Optimal Cutting Scores in Criterion-Referenced Measurement*. Journal of Experimental Education, 45, pp.4-9.
- Jager, M. (Summer 1991) *Selection of Judges for Standard-Setting*. Educational Measurement: Issues and Practice, 10, pp.3-6. 10, 14.
- Livingston, S.A., and Zieky, M.J. (1982) *Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests*. Educational Testing Service, Princeton, New Jersey.
- Mills, C.N., Melican, G.J. and Ahluwalia, N.T. (Summer 1991) *Defining Minimal Competence*, Educational Measurement: Issues and Practice, 10, pp. 7-9.
- Plake, B.S. (Summer 1991) *Factors Influencing Intrajudge Consistency During Standard-Setting*. Educational measurement: Issues and Practice, 10, p. 15-16, 22, 25-26.
- Reid, J.B. (Summer 1991) *Training Judges to Generate Standard-Setting Data*. Educational Measurement: Issues and Practice, 10, pp. 11-14.
- Shepard, L. (1983) *Standards for Placement and Certification*. In Anderson, S.A., and Helmick, J.S. (Eds) *On Educational Testing*. Washington, D.C.: Jossy-Bass, Chapter 4.

---

**Additional  
information  
(Continued)**

Smith, R.L. and Smith, J.K. (1988) *Differential Use of Item information by Judges Using Angoff and Nedelsky Procedures*. Journal of Educational Measurement, 25, pp. 259-274.

van der Linden, W. (1982) *A Latent Trait Method for Determining Intrajudge Inconsistency in the Angoff and Nedelsky Techniques of Standard-Setting*. Journal of Educational Measurement, 4, pp. 295-305.

---

## Section B

### Criterion-Referenced Test Analysis

---

#### Overview of norm-referenced and criterion-referenced test analysis

There are a number of statistics available for Criterion-Referenced (CR) test analysis. The use of Norm-Referenced (NR) statistics to analyze CR tests is questionable. This section will discuss NR and CR testing, the use of CR statistics for test analysis, and some lessons learned for CR test analysis.

Until the 1970's, most educational measurement was *Norm-Referenced*. Instructional programs and schools were not sensitive to individual needs or performance characteristics except in terms of *how the individual performed in relation to the individuals' peer group*.

To assist educators in measuring the effectiveness of their tests, psychometricians developed an array of statistics for measuring test reliability on *statistically normal population distributions*, hence the name Norm-Referenced.

Companies (e.g., Educational Testing Service, and American College Testing) specialized in the development of so-called, standardized (Norm-Referenced) tests, and developed powerful statistical procedures to detect subtle weaknesses in their tests and enable them to apply test results to larger and larger groups of individuals.

---

#### Arguments against using NR statistics for small-groups

One of the major reasons for not using NR measurement for classroom or small-group instruction is that *a single classroom or small group of individual students (i.e., 20-30) very likely does not represent a statistically normal distribution, regardless of the attribute being measured*.

This fundamental assumption of NR statistics is violated in small-group test analysis.

NR statistics computed on data from a *non-normal distribution* cannot be meaningfully or reliably interpreted.

---

#### Arguments for using NR statistics for small-groups

The assumptions on which NR statistics are based are robust. Any assumptive violations are minor.

As group size increases to several hundred or more students, the distribution of the attribute being measured tends to become normally distributed.

---

---

**Central limit theorem**

This Central Limit Theorem allows the use of NR statistics on statistically non-normal groups with some assurance that interpretations on *aggregated data* will be meaningful.

The use of NR statistics is justified for examining test effectiveness on the basis of data gathered across large numbers of students.

NR statistics are conceptually and mathematically less complex than CR statistics.

---

**Advantages of NR test analysis over CR test analysis**

NR test analysis has distinct advantages over CR test analysis that makes it superior in certain test analysis situations:

NR test analysis is capable of producing information from which individuals may be *ranked*. When it is necessary to separate or rank-order individuals for any reason, NR statistics are the statistics of choice.

*NR statistics are parametric in nature*, since they are based on assumptions about the statistical parameters of the distribution being measured. *CR statistics are, by definition, nonparametric* since they make no assumptions about the underlying population parameters.

In general, parametric statistics have greater statistical power than do nonparametric statistics.

---

**Why not use NR statistics for CR test analysis?**

The primary reason not to use NR statistics for CR test analysis is that *achievement scores in CR instructional situations are not expected to be normally distributed*. Therefore, problems can arise when evaluators attempt to use NR test analysis statistics to measure the effectiveness of CR-type instruction.

The primary interest in CR test analysis is to determine *to what extent the students achieve the instructional objectives of the course, not how a student performs in relation to other students*.

The focus of CR instruction is the *mastery of objectives*, and the mastery of the course content.

In CR instruction, the student evaluation emphasis is on the percentage of course material that each student has mastered at any point in the course, and *not on the student's relative achievement* with respect to other students in the class or any other norming group.

---

---

**Differences in NR and CR tests**

In a NR environment, the test needs to reliably (consistently) measure student performance (achievement) in relation to other students.

NR test analysis statistics indicate how far a student's score deviates from the group mean.

In a CR environment, the test must reliably (consistently) measure whether or not a student has reached a particular level of mastery in relation to an absolute (non-relative) cut-off score (criterion).

CR test analysis statistics indicate how far a student's score deviates from a fixed standard, the criterion.

---

**Some differences in NR and CR test analysis statistics**

The statistics used to analyze CR and NR tests are different because the purposes of CR and NR measurement are different.

CR measurement provides additional meaning from test scores by referencing the test outcome to a clearly-specified body of test content.

NR measurement provides meaning from test scores by comparing the test outcome to other scores.

CR instructional programs force test score distributions to be non-normal. Therefore, NR correlation coefficients (such as Kuder-Richardson 20 and 21) may be inadequate reliability indices for CR tests.

CR reliability coefficients (e.g., Livingston's  $k^2$ ) measure reliability (repeatability) by assessing deviations from the criterion score. (NR reliability coefficients measure reliability (repeatability) by assessing deviations from the group mean.)

CR testing theory encompasses "domain theory". Domain theory includes the issues of appropriate sampling and representativeness of test items. These issues are typically not addressed in NR testing theory.

The fact that an evaluator never knows the true score of any individual in terms of the knowledge domain being tested requires the use of sophisticated CR test statistics.

---

---

**Test statistics:  
Indices of CR test  
reliability**

Indices of CR test *reliability*, (also called indices of CR test *agreement*) fall into three categories:

- Threshold Loss.
- Squared-Error Loss.
- Domain Score Estimation.

The particular category selected by an evaluator depends upon the interpretations of the statistic that are desired.

---

**Use of threshold  
loss agreement  
indices**

*Threshold loss agreement indices* are used to measure a test's ability to consistently classify masters and non-masters correctly, especially when misclassification errors are equally serious regardless of how large they are (i.e., misclassification of students whose actual scores are close to the cut-off (criterion) score is just as serious as misclassification of students whose actual scores are far from the cut-off (criterion) score).

---

**Use of squared-  
error loss  
agreement indices**

*Squared-error loss agreement indices* deal with test score consistency as related to *degrees of mastery* rather than simply classification. Squared-error loss agreement indices are used when the consequences of misclassifying students who are far above or below the cut-off (criterion) score are considered more serious than the consequences of misclassifying those students who are close to the cut-off (criterion score).

---

**Use of domain  
score estimation  
agreement indices**

*Domain score estimation agreement indices* are not concerned with issues of classification of students, but instead are designed to estimate the stability of test scores or the proportion of items from a given domain that students answer correctly.

---

**Statistics for  
threshold loss  
agreement indices**

There are two threshold loss agreement statistics, the *Agreement Index* ( $P_o$ ), and the *Kappa Coefficient* ( $k$ ). Both of these statistics are based on two fundamental assumptions:

- Student classifications are dichotomous (masters or non-masters) based on a cutoff (threshold or criterion) score.
-

**Statistics for  
threshold loss  
agreement indices  
(Continued)**

Student misclassification errors are equally serious whether student scores are close to, or far from, the cutoff score.

*The  $P_o$  agreement index is designed to be used with two administrations of the same, or parallel test form (e.g., a pre-test and a post-test).*

**Computation of the  
agreement index  
( $P_o$ )**

*Example of Computation of the Agreement Index ( $P_o$ )*

In this example, 20 students were administered two parallel tests (or the same test twice). The cut-off criterion score was 18 on both tests.

One student obtained a mastery score (>18) on both tests. Therefore, a 1 is entered into the following table in the mastery-mastery cell.

One student obtained a mastery score on the first test but not on the second test. Therefore, a (1) is entered into the table in the mastery-nonmastery cell.

No students were below the cutoff score on the first test and above the cutoff score on the second test. Therefore, no students were classified as nonmastery-mastery.

The remaining 18 students did not obtain a mastery score on either test. Therefore, 18 is entered into the table in the nonmastery-nonmastery cell.

<b>Test Number 2</b>			
<b>Test Number 1</b>	Mastery	Nonmastery	Total
Mastery	1	1	2
Nonmastery	0	18	18
Total	1	19	

To compute  $P_o$  add the proportion of students classified as masters on both tests and the proportion classified as non-masters on both tests:

$$\begin{aligned} P_o &= 1/20 + 18/20 \\ &= 19/20 \\ &= 0.95 \end{aligned}$$

If all students were consistently classified as masters or non-masters (not the case in this example),  $P_o$  would be 1.00, indicating perfect classification. This is the upper limit of  $P_o$ .

---

**Computation of the agreement index ( $P_o$ ) (Continued)**

The lower limit of  $P_o$  is the proportion of consistent classifications on two tests expected by chance alone. To compute the chance proportion ( $P_c$ ), the margin totals in the table are used as a percentage of the total number of students tested:

$$\begin{aligned} P_c &= 2/20 \times 1/20 + 18/20 \times 19/20 \\ &= 2/400 + 342/400 \\ &= 0.86 \end{aligned}$$


---

**Computation of Kappa coefficient**

Example of Computation of the Kappa Coefficient ( $k$ )

Using the  $P_o$  and  $P_c$  statistics,  $k$  is computed as follows:

$$\begin{aligned} k &= (P_o - P_c) / (1 - P_c) \\ &= (.95 - .86) / (1 - .86) \\ &= 0.64 \end{aligned}$$

These computations for  $P_o$  and  $k$  are for two administrations of the same or parallel tests (e.g., pre-test-post-test situations). Methods for computing  $P_o$  and  $k$  for single test administration have been developed, but the methods are computationally complex and usually require computer support.

---

**Squared-error loss agreement indices:  $k^2$  Index and  $\Phi(\lambda)$  index**

There are two agreement indices in this category:

The  $k^2$  Index, which is the ratio of true test scores to squared deviations of expected scores from the mean.

The  $k^2$  Index assumes that a classically parallel test form is used for student classification.

The  $\Phi(\lambda)$  Index, which is a measure of the dependability of a test to correctly classify individuals.

The  $\Phi(\lambda)$  Index assumes that a randomly parallel test form is used for student classification.

---

---

**Classically parallel tests**

Classically parallel tests are generated by developing two or more samples of questions that are as comparable as possible, are based on the same set of course objectives, and are from the same item domain.

When properly developed, classically parallel tests will yield the same mean, variance, and intercorrelational statistics, making them parallel.

---

**Randomly parallel tests**

Randomly parallel tests are generated by selecting simple random or stratified samples of items from an item domain.

Randomly parallel tests are unlikely to yield the same mean, variance, and intercorrelational statistics.

One of the advantages of the  $\Phi (\lambda)$  Index is its ability to be used to measure the dependability of randomly parallel tests, such as computer-generated tests.

---

**Statistics for domain score estimation**

Domain Score Estimation statistics are concerned with estimating the stability of an individual's score or proportion correct in an item domain, independent of any mastery standard.

Domain Score Estimation statistics are useful for estimating a confidence interval around the cutoff score for determining the accuracy of mastery-nonmastery decisions from a single test administration.

Domain Score Estimation statistics require only a single test administration, and are readily calculated by hand.

Domain Score Estimation statistics are not designed for classifying students as masters or non-masters.

---

**Test item statistics:  
Norm-Referenced**

*Norm-Referenced (NR) test makers are concerned about test items in terms of:*

Their level of difficulty.

Their ability to discriminate between high and low achievers.

The distribution of students among the various response alternatives.

---

---

**Test item statistics:** *Criterion-referenced (CR) test makers are concerned about test items in terms of:*  
**Criterion-referenced**

Their level of difficulty.  
Their ability to discriminate between high and low achievers.  
*The major concern is the test item's performance with uninstructed versus instructed (non-masters versus masters)*

---

**Test item difficulty** The concept of item difficulty is fundamentally the same in either the NR or CR testing environment: the number of students getting the item correct divided by the total number of students tested times 100 (to convert the quotient to a percentage).

---

**Difficulty index for norm-referenced tests**

*Determining the Difficulty Index for NR Test Items*

For NR tests, it is sufficient to determine the difficulty index for test items administered only to the *instructed group* of students (students who have received instruction). Test items administered to students prior to instruction are assumed to have high difficulty indices, therefore item difficulty analysis is not typically done on NR pre-tests.

---

**Difficulty index for criterion-referenced tests**

*Determining the Difficulty Index for CR Test Items*

In the CR testing environment, it is important to determine the difficulty index of items on both the pre-test and post-test (to both uninstructed and instructed groups).

If difficulty levels are determined only for instructed students, there is no way of knowing whether students would have performed similarly without the instruction.

---

**CR test item  
discrimination  
statistics**

*Pre-test-Post-test Difference (DIS<sub>PPD</sub>)*

This CR test item difficulty level statistic ranges in value from -1.00 to +1.00. The advantage of this statistic is that it is not sensitive to individual performance changes, only group gains or losses.

To compute the DIS<sub>PPD</sub>, the proportion of students answering an item correctly on the pre-test is designated P<sub>pre</sub>. The proportion of students answering an item correctly on the post-test is designated P<sub>post</sub>.

$$DIS_{PPD} = P_{post} - P_{pre} \text{ for each test item.}$$

**CR test item  
difficulty:  
Individual gain**

*Individual Gain (DIS<sub>IG</sub>)*

This CR test item difficulty level statistic ranges in value from 0 to +1.00, and measures the proportion of students who actually gained from instruction. It is computed as the proportion of students who answered a test item incorrectly on the pre-test (Q<sub>pre</sub>) and correctly on the post-test (P<sub>post</sub>). The proportion of students answering an item correctly on the pre-test is designated P<sub>pre</sub>.

$$DIS_{IG} = Q_{pre} / P_{post} = (1 - P_{pre}) / P_{post}$$

DIS<sub>IG</sub> produces an inflated measure of instructional effectiveness because it does not consider the negative effect of students who answered incorrectly on both the pre-test and the post-test (those students who evidently did not learn from the instruction).

**CR test item  
difficulty: Net gain**

*Net Gain (DIS<sub>NG</sub>) External Sensitivity Index*

The Net Gain statistic corrects for the inflation in instructional effectiveness generated by the Individual Gain statistic by subtracting from DIS<sub>IG</sub> the proportion of students who answered the item incorrectly on both the pre-test and the post-test.

$$DIS_{NG} = DIS_{IG} - (Q_{pre} / Q_{post})$$

**CR test item  
difficulty: Net gain**

The Net Gain statistic ranges in value from -1.00 to +1.00. Essentially, it considers only those students who answered the test item incorrectly on the pre-test (the only students who could possibly gain from instruction).

**CR test item  
difficulty:  
Unobstructed  
group differences***Uninstructed-Instructed Group Difference* ( $DIS_{UIGD}$ )

A student must have completed instruction and have been administered the post-test before the Pre-test-Post-test Difference, Individual Gain, or Net Gain (External Sensitivity Index) statistics can be computed.

The Uninstructed-Instructed Group Difference statistic has the advantage of being computed at the beginning of instruction. The  $DIS_{UIGD}$  requires that two groups be tested--a group already possessing mastery in the content (instructed group), and a group possessing no mastery (uninstructed group).

The  $DIS_{UIGD}$  statistic is computed by subtracting the proportion of students in the instructed group (denoted as  $P_{mastery}$ ) who answered the item correctly from the proportion of students in the uninstructed group (denoted as  $P_{nonmastery}$ ) who answered it correctly.

$$\begin{aligned} DIS_{UIGD} &= P_{mastery} - P_{nonmastery} \\ &= U/N1 - L/N2 \end{aligned}$$

where:

U is the number of students at or above the cutoff score who answered the item correctly,

L is the number of students below the cutoff score who answered the item correctly,

N1 is the total number of students at or above cutoff, and

N2 is the total number of students below cutoff.

The  $DIS_{UIGD}$  statistic ranges in value from -1.00 to +1.00.  $DIS_{UIGD}$  uses separate cutoff scores for the uninstructed and instructed groups.

---

**Lessons learned  
for CR test analysis**

Consider the following:

A number of statistics are available for use in CR testing situations.

Some of them are interpretationally analogous to more familiar statistics used in NR testing environments.

The majority of CR test statistics are not difficult to compute and interpret. The CR test statistics discussed in this handbook are easily computed by hand or are readily estimated from a table. Their interpretations are relatively straightforward and meaningful.

*Norm-referenced test statistics* do not provide sufficient information for making decisions regarding proficiency or mastery of a domain of content from test results of a sample of content from the domain.

*Criterion-referenced test statistics* are made for the purpose of providing sufficient information for making decisions regarding proficiency or mastery of a domain of content from test results of a sample of content from the domain.

---

**Additional  
information**

The following bibliography provides additional information on CR test analysis:

Berk, R.A. (1984) *A guide to Criterion-Referenced test construction*. Baltimore, MD: Johns Hopkins University Press, Chapters 5 and 9.

Brennan, R.L. (1972) A generalized upper-lower item Discrimination Index. *Educational and Psychological Measurement*, 32, pp. 289-303.

Brennan, R.L. and Kane, M.T. (1977) An index of dependability for mastery tests. *Journal of Educational Measurement*, 14, pp. 277-289.

Cohen, J. (1977) *Statistical power analysis for the behavioral sciences*. (Rev) New York: Academic Press.

Cox, R.C. and Vargas, J.S. (February 1966) *A comparison of item selection techniques for Norm-Referenced and Criterion-Referenced tests*. Paper Presented at the National Council on Measurement in Education Conference, Chicago, IL

---

---

**Additional  
information  
(Continued)**

- Kane, M.T. and Brennan, R.L. (1980) Agreement coefficients as indices of dependability for domain-referenced tests. *Applied Psychological Measurement*, 4, pp. 105-126.
- Livingston, S.A. (1972) Criterion-referenced applications of classical test theory. *Journal of Educational Measurement*, 9, pp. 13-26.
- Millman, J. (1980) Computer-based item generation. In R.A. Berk (Ed.), *Criterion-referenced measurement: The state of the art*. Baltimore, MD: Johns Hopkins University Press, p. 32.
- Shavelson, R.J., Block, J.H. and Ravitch, M.M. (1972) Criterion-referenced testing: Comments on reliability. *Journal of Educational Measurement* 9, pp. 133-137.
- Subkoviak, M.J. (1976) Estimating reliability from a single administration of a mastery test. *Journal of Educational Measurement*, 13, pp. 265-276.
- Subkoviak, M.J. (1980) Decision-consistency approaches. In R.A. Berk (Ed.), *Criterion-referenced measurement: The state of the art*. Baltimore, MD: Johns Hopkins University Press, p. 129-185.
- Subkoviak, M.J. (1984) Estimating the reliability of mastery-nonmastery classifications. In R.A. Berk (Ed.), *Criterion-referenced measurement: The state of the art*. Baltimore, MD: Johns Hopkins University Press, p. 32.
- Subkoviak, M.J. (1988) A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement*, 25, pp. 47-55.
-

## Section C

### Guidelines For Test Validation

---

#### Common types of test validity: Face validity

There are five common types of test validity.

#### Face Validity

Face validity is a measure of what a test appears to measure, not what is actually measured.

appears to cover the content of a course of instruction.

This subjectivity makes face validity the weakest form of test validity because it involves the least amount of rigor in its assertion.

The validity of a test instrument should not be determined based solely on face validity.

Face validity is also known as *prima facie* validity. It is not considered a formal type of test validity because of its inherent weaknesses.

---

#### Common types of test validity: Content validity

#### Content Validity

Content validity is a measure of how closely the test instrument relates to the content of the instructional program it is designed to measure.

Both the test questions and the course content should be directly associated with specific instructional objectives.

The content domain should be systematically sampled using formal Task Analysis methodology.

The content domain tasks and subtasks identified through the task analysis process should form the basis for instructional objectives.

---

**Common types of  
test validity:  
Content validity  
(Continued)**

The degree of content validity is not expressed numerically. It is described in terms of the comparison between, or the correspondence among,  
     course objectives,  
     course content, and  
     test questions.

Evidence of the degree to which a test is deemed to be content valid is based on the combination of

    The comparison of test questions with course content,  
     and

    The comparison of test questions and course content with instructional objectives.

**Common types of  
test validity:  
Predictive validity**

Predictive Validity (Criterion Validity)

Predictive validity is a measure of how well predictions made by a test are confirmed by some future behavior of students or graduates of the instructional program. (The ability of a test to predict criterion performance behavior)

Predictive validity of multiple variables can be determined using a mathematical model known as a regression equation. The output of the regression equation using more than one predictor variable is called a multiple regression coefficient (R). The R statistic is a measure of prediction of job performance measures from more than one predictor. Predictive validity is determined by:

    Administering the test,

    Waiting for a time until the behavior (criterion variable) of interest is displayed by the student or graduate of the instructional program, and

    Correlating the individual's test score with the level of performance of the behavior.

**Common types of  
test validity:  
Predictive validity  
(Continued)**

Direct correlation occurs when high test scores relate to high performance scores (or low test scores relate to low performance scores).

Indirect correlation occurs when high test scores relate to low performance scores (or low test scores relate to high performance scores).

**Common types of  
test validity:  
Concurrent validity**

Concurrent Validity

Concurrent validity is closely related to predictive validity, and is a measure of whether two different tests are related closely enough to each other that they both could be considered to be measuring the same attributes.

The distinction between concurrent and predictive validity depends on whether the two tests are administered at the same time or are separated by a period of time. If the tests are administered together, the measurement of validity is concurrent.

Tests with high concurrent validity can often be interchanged if they can be shown to measure the same attributes (criteria).

Two tests can correlate highly with one another (high concurrent validity), but one or both of the tests may not be highly correlated with the real-life criterion behavior.

Do not assume that two different tests each measure what they are supposed to measure, even if they have high concurrent validity.

Content validity is a prerequisite for predictive and concurrent validity.

**Common types of  
test validity:  
Construct validity**

Construct Validity

Construct validity is the degree to which scores on a test permit inference about underlying traits.

Identifying traits that legitimately comprise the hypothetical construct.

Finding/devising instruments able to accurately measure the identified traits.

Predictive and/or construct validity is often a prerequisite for concurrent validity.

---

**Predictive criterion  
and construct  
validation of test  
instruments**

*Content and predictive (criterion) validity* must relate to some underlying constructs within the knowledge domain and/or the real world. These constructs serve as anchors for test items.

Typically, *instructional objectives are used as construct anchors to tie test questions to course content*. In the formal Instructional System Development process, both instructional objectives and test items to measure student mastery of the objectives are developed immediately after the task or job analysis phase.

The entire *task analysis process provides the basis* for the development and either predictive (criterion) or construct validation of the test instruments in the course of instruction.

In most Air Force instructional systems, instructional objectives define either intellectual skills or psychomotor behavioral skills. Since these criterion behaviors are observable, *predictive (criterion) validity is the most important kind of validity for Air Force Criterion-Referenced tests*.

Some psychologists view construct validity as the most important kind of validity due to the contention that intellectual skills or psychomotor behavioral skills are related to some underlying traits. Therefore, construct validity is seen as the “umbrella” concept for validation, and content, concurrent, and predictive validity are special cases of construct validity.

---

---

**Unique examples  
of validation**

*Content validation* is the most common kind of test validity used by military organizations. *Content validity* is a measure of how closely the test instruments relate to the content of the instructional program it is designed to measure. Some organizations have used unique techniques to determine content validity.

The Extension Course Institute (ECI) used the technique of *factor analysis* to determine the underlying structure of survey instruments administered following courses of instruction. Factor analysis revealed question sets that, when pooled together, described particular attributes and attitudes of students in the course, and suggested likely areas where the survey instrument could be restructured.

Identified clusters of questions that seemed to fit together even though the questions were placed in different sections of the survey.

Demonstrated that factor analysis may be capable of validating the internal structure of a measurement instrument as well as identifying links (anchors) between the measurement items and external constructs.

*Predictive validity* is a measure of how well predictions made by a test are confirmed (correlated) by some future behaviors of students or graduates if the instructional program. *Concurrent validity* is closely related to predictive validity, and is a measure of whether two different tests are related closely enough to each other that they both could be considered to be measuring the same attributes. Some organizations have used unique techniques to determine *predictive and concurrent validity*.

The Educational Testing Service (ETS) uses *correlation analysis* in a variation of predictive and concurrent validation techniques to analyze tests.

Graduates of a course of instruction and their supervisors are surveyed some time after completion of the course. The survey asks questions dealing with the individual's knowledge and abilities in areas that relate to the course of instruction.

---

---

**Unique examples  
of validation  
(Continued)**

Supervisor and graduate surveys are compared. Pairs of survey instruments with more than 85% agreement are subjected to correlation analysis.

Graduate scores on their course tests are correlated against the judgments expressed in the survey. High correlations suggest that the *courses are teaching what is needed on the job.*

---

**The need for  
construct  
validation in  
military institutions**

*Construct validity* is the degree to which scores on a test permit inference about underlying traits. *Without construct validation, it is impossible to determine why scores on a test vary.*

It is generally assumed that variance in test scores is due to differences in student knowledge of the subject or differences in performance ability. This may be true, but proof is required to determine if the test measures only one concept, or if the test measures two different concepts simultaneously.

For example, a performance test in electronic circuit analysis may measure not only a student's cognitive knowledge of electronics, but also mathematical ability (skill).

Without performing construct validation on this test, it will be difficult to determine *what percentage of the test score variance* is attributable to electronics *knowledge* and what percentage is attributable to mathematical *skills*.

Without detailed knowledge of variance distribution in a test, it is difficult to improve the effectiveness of the test.

There are several other issues to consider concerning the need for construct validation of tests.

Construct validation of a test also validates the *use of the test*.

Construct validation answers the question of *what is the proper purpose of a test?*

Construct validation can determine the *degree of representativeness of a test for measuring a domain of knowledge*.

A test represents a domain if the domain of knowledge was adequately sampled to ensure that the test represents a certain depth or breadth of coverage of the domain.

A test represents a domain if inferences can be drawn from the test items back to the domain.

---

---

**Additional  
information**

The following bibliography provides additional information on test validation:

- Anderson, Scarvia B. In Anderson, S.B., Ball, S., and Murphy, R.T. (1975). *Encyclopedia of Educational Evaluation*, San Francisco: Jossey-Bass, Inc.
- Borg, W.R. and M.D. Gall (1979). *Educational Research: An Introduction* (Third Edition), New York: Longman.
- Diehl, G.E. (1989). *Factor Analysis of the ECI Course for Authors- Graduate Survey*, July, 1989.
- Jensen, A.R. (1980). *Bias In Mental Testing*, New York: The Free Press.
- Messick, S. *The Meaning and Consequences of Measurement*, in Wainer, H. and Brown, H.I. (Eds) (1988). *Test Validity*, Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc.
- Rubin, D. In Wainer, H. and Brown, H.I. (Eds) (1988). *Test Validity*, Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc.
-

## **Section D**

### **Lessons Learned for Test Item Development**

---

**Introduction**

The use of the following methodologies for test and measurement analysis, development and administration have become more common:

- Criterion-referenced test analysis.
- Criterion-referenced test statistics.
- Item Response Theory (IRT).
- Computer Adaptive Testing (CAT).

The development of Criterion-Referenced test statistics for test analysis, and the methodology for Criterion-Referenced standard-setting is not new, however, the computational hardware and software has recently become accessible to test and measurement practitioners to use these methodologies effectively.

Computer technology has also enabled the practical application of Item Response Theory (IRT) and Computer Adaptive Testing (CAT) that is based upon IRT concepts. In the past, the time delay for selection and display of CAT test items significantly detracted from the instructional effectiveness of CAT. Current computers running at high megahertz rates and incorporating math coprocessors are capable of reducing the computational delays to acceptable levels.

---

**This section provides**

This section provides a discussion of:

- Lessons learned for test item development, including test formats and design, and test scoring methods.
- Lessons learned for test and item analysis.
- Lessons learned for test administration.
- Summary of lessons learned for test item development.
- Lessons learned for establishing local norms for test statistics.

---

---

**Test formats and design: Alternative choice (AC) test items****Alternate-Choice (AC) Test Items**

The Alternate-Choice test item format was originally proposed by Ebel in 1982 to overcome some of the difficulties with traditional true-false test items.

Characteristics of Alternative-Choice test items include:

The stems are usually short.  
Two options are provided.  
Options generally consist of very short phrases or single words.

The following is an example of an AC test item:

According to Ohm's Law, to maintain a constant voltage in a circuit, the current flow would have to be (1) doubled (2) cut in half if the resistance in the circuit were doubled.

---

**Benefits of AC test items**

Benefits of AC test items compared to true-false test items are:

AC test items are less ambiguous than true-false test items. True-false test items require the examinee to make comparisons between two choices rather than requiring an absolute judgment about the truth or falsity of the premise. AC test items are suitable for testing the lower cognitive levels of learning by increasing the sophistication of the question. The levels of learning defined in Bloom's taxonomy are:

- Knowledge
- Comprehension
- Application
- Analysis

AC test items are not suitable for testing Bloom's higher levels of learning (Evaluation and Synthesis).

AC test items also are suitable for testing the lower-level intellectual skills defined in AFM 36-2234, Instructional System Development:

- Discrimination
- Concrete Concepts
- Defined Concepts
- Rule Learning

---

**Disadvantages of AC test items**

AC test items are not suitable for testing the higher-level *intellectual* and *motor skills* defined in AFM 36-2234:

- Verbal Information
- Cognitive Strategies
- Metacognition
- Motor Skills
- Attitudes and Motivation

AC test items can easily be developed to measure cognitive levels of learning (intellectual skills) above the factual level of learning, which is what most true-false test items measure.

**Benefits of essay test items***Essay Test Items*

Benefits of essay test items include the following:

- More appropriate for measuring critical thinking, analysis, and logical organization than any other test item form.
- Can test the highest levels of learning such as Synthesis and Evaluation (Bloom's Taxonomy) or Cognitive Strategies (AFM 36-2234) better than any other test form.
- Relatively easy and quick to develop.
- Complexity can be easily adjusted to match the presumed (testable) level of complexity in student's thinking according to such factors as age, ability, and experience.

**Disadvantages of essay test items**

Disadvantages of essay test items include the following:

- Difficulty, costs and unreliability of scoring essay items.
- Inherent limited sampling of a content domain.
- Especially difficult to use in distance learning environments.
  - Expert faculty is required to grade essay items.
  - High student-to-faculty ratios exacerbate delays in scoring and processing tests.
  - Essay tests may be perceived as unfair, unreliable, or invalid.
  - Additional resident faculty or adjunct faculty members may be required to develop and score essay questions.

---

**Essay scoring reliability**

Recommendations to lessen the problems associated with the scoring reliability of essay tests include:

The easiest way to increase the reliability of any test is to increase the number of test items.  
Improve the reliability of essay tests by restricting the length of the answers required of students. This will increase the number of possible questions asked in a given amount of time. Increasing the number of questions also improves the test's ability to more representatively sample the content domain.  
To the maximum extent possible, *develop key word and phrase metrics to generate a numerical score for essay tests. Criteria for number of key words and phrases correct can be established to generate a numerical passing score.*  
If possible, *develop time criteria for answering test items.*  
These metrics can be combined with the key word and phrase metrics to generate *time and error criteria for the test.*

---

**Advantages of true-false test items***True-False Test Items*

Advantages of true-false test items:

Easiest of all test forms to write and administer.  
Possible to test a large domain of knowledge in a relatively short time period.  
Easy to grade.

---

**Disadvantages of true-false test items**

Disadvantages of true-false test items:

Lower reliability than multiple-choice test items.  
True-false items are easier than multiple-choice items from a student's perspective. True-false test items have only two alternatives to choose from.  
True-false test items tend to be less discriminating than multiple choice test items due to a lower variability in response options.

---

---

**Ways to improve true-false test items**

Some ways to improve true-false test instruments are:

Increase reliability by increasing the number of items per test.  
Improve discrimination by ensuring that the questions are based on course objectives, rather than on arbitrarily selected sentences out of the course textbooks or other documentation.

---

**Multiple-choice test items***Multiple-Choice Test Items*

Multiple-choice are the most popular type of test item.

Critics of multiple-choice test items contend that they measure only lower-order intellectual skills (recall of factual information such as discriminations, concepts, and rule using) rather than testing higher-order intellectual skills.

Multiple-choice test items can be written to measure high-order intellectual skills such as application, analysis, synthesis, and evaluation (Bloom's Taxonomy), or cognitive strategies (AFM 36-2234).

---

**Short answer and completion test items**

Short Answer and Completion test items can be thought of as hybrids of alternate choice, true-false, multiple choice, and essay test items.

The formats are similar to essay test items in that they permit some form of creative, free response.

The formats are similar to multiple-choice test items in that examinee responses are limited in form and scope by the context of the question in which they are embedded.

The formats do not allow for the full range of free response permitted in the essay test item.

The formats are not as easily or objectively measured as true-false or multiple-choice test items.

---

**Matching test items** Matching test items can be thought of as extended multiple-choice test items.

Matching test items are better suited for knowledge-level (factual and discrimination) and comprehension-level (concept and rule using) intellectual skills.

The major drawback of matching test items is the difficulty of measuring the highest-order intellectual skills such as synthesis and evaluation (Bloom’s Taxonomy) or cognitive strategies (AFM 36-2234).

**Test item / knowledge level grid**

The following table is a Knowledge-level/Type of Test Item Grid for Bloom’s Taxonomic levels of learning. Within each cell of the grid, a (++) indicates the particular test item is ideal for testing the particular level of learning. A (+) indicates the test item is satisfactory for testing the level of learning. An empty cell indicates the test item is not acceptable for testing that particular level of learning.

Table 10 Types of Test Items for Bloom’s Levels of Learning

	True-False	Short Ans./Completion	Matching	Multiple-Choice	Essay
Evaluation				+	++
Synthesis			+	++	++
Analysis		+	+	++	++
Application		+	++	++	++
Comprehension	+	++	++	++	+
Knowledge	++	++	++	++	+

**Intellectual skill-level / test item grid**

The following table is a Intellectual Skill -level/Type of Test Item Grid for the Intellectual Skills in AFM 36-2234. Within each cell of the grid, a (++) indicates the particular test item is ideal for testing the particular level of learning. A (+) indicates the test item is satisfactory for testing the level of learning. An empty cell indicates the test item is not acceptable for testing that particular level of learning.

**Intellectual Skill-  
Level / Test Item  
Grid (Continued)**

Table 11 Types of Test Items for AFM 36-2234 Intellectual Skills

	True-False	Short Ans./Completion	Matching	Multiple-Choice	Essay
Cognitive Strategies		+	+	++	++
Rule-Learning		+	++	++	++
Concrete Concepts	+	++	++	++	+
Defined Concepts	+	++	++	++	+
Discrimination	++	++	++	++	+

**Test scoring  
methods:  
Traditional**

Using traditional test scoring methods, each test item is assigned one point, which is awarded to the examinees that select the most correct answer. Examinees who do not select the most correct answer receive zero points.

This traditional scoring method is intuitive in use and interpretation. It is also computationally straightforward.

**Non-traditional test  
scoring methods:  
Admissible  
probability testing**

Other scoring mechanisms have been developed to serve particular purposes. This section will discuss three non-traditional scoring methods.

*Admissible Probability Testing (APT)*

APT is based on the fact that examinees who miss an item on a test probably still know something about the test item (other than the correct answer).

Traditional scoring methods give zero credit to the examinee who misses a test item. Awarding zero credit is actually a statement that the examinee does not possess the knowledge in the area tested.

In reality, zero credit is warranted only if the examinee truly knows nothing about the test item. Very often, examinees do know something about a test item, even if they do not know the correct answer. Traditional scoring does not give examinees credit for such knowledge.

**Non-traditional test scoring methods:  
Admissible probability testing  
(Continued)**

APT gives credit for this kind of knowledge by allowing the examinee to assign *probabilities of correctness to each alternative of a multiple-choice, alternate-choice, or matching test item.*

Since the concept of probabilities is foreign to many examinees, the instructions for a *four-alternative multiple-choice test item*, for example, are for the examinees to assume they have \$100 to “spend” on their answer to the test item. For example:

Who was the 16th President of the United States?

- a. Grover Cleveland
- b. Abraham Lincoln (correct answer)
- c. Thomas Jefferson
- d. Dwight Eisenhower

Examinees may put as much as they want on any single answer or combination of answers as long as they “spend” the entire \$100.

For example, an examinee who is sure that the correct answer is (b.) may elect to place the entire \$100 on answer (b.). Another examinee may not be as sure of the correct answer and place \$50 on each of two alternatives. Another examinee may not know anything about the test item and place \$25 on each of the four alternatives.

In effect, the examinee is estimating the probability of being correct at 1.00 (for absolute certainty); 0.50 probability if the \$100 is split between two alternatives; or 0.25 probability if the \$100 is distributed equally across all four choices.

In the four-alternative multiple-choice test item example, the scoring algorithm for APT is designed to give full credit to the examinee who puts the entire \$100 on the correct answer (1.0 probability); the next highest credit to the examinee who splits the \$100 across two alternatives, one of which is correct (0.50 probability); and the least credit to the examinee who splits the \$100 across all four alternatives (0.25 probability).

The APT scoring method has several disadvantages which explain why the scoring technique is not widely used:

---

---

**Non-traditional test scoring methods:  
Admissible probability testing  
(Continued)**

The APT scoring algorithm is computationally complex, and requires a computer to generate test item and test scores. There is actually a reward for guessing, once the examinee figures out the scoring scheme (“spending” equally on each alternative).

The APT scoring algorithm ensures that an examinee can not pass a test by guessing and “spending” equally across all test item alternatives. However, the total point score on a test may give a false impression of an examinee’s ability by inflating the total score.

---

**Non-traditional test scoring methods:  
Weighted alternative scoring**

*Weighted Alternative Scoring*

Multiple-choice tests can be used to test higher-order cognitive skills. *There are difficulties in writing multiple-choice test items that measure these higher-order skills:*

Writing multiple-choice test items that measure an individual’s ability to *synthesize concepts into a larger whole, or to evaluate the relative value of a situation* are difficult to construct because the line between correct and incorrect become more difficult to define.

It is difficult to invent plausible yet incorrect distracters for multiple-choice test items that measure higher-order cognitive skills.

The difficulty becomes even more complicated when the topic being tested is a “soft skill” such as leadership, management, human relations, etc.

---

**Differentially weighted alternatives**

The technique of *differentially weighting alternatives* can be used to alleviate the difficulties of developing multiple-choice test items to measure higher-level cognitive skills.

Complex “soft skill” topics seldom have clearly right or wrong answers.

To structure tests that have clearly measurable right or wrong answers would result in superficial assessment of student’s abilities using artificial, hypothetical situations.

Test items can be written with *one most correct answer*. This answer is assigned maximum credit (5 points, for example).

---

---

**Differentially weighted alternatives (Continued)**

This *test item weight* is used as the *maximum weight for that test item*.

Two or three *distracters* are written which are *partially correct in varying degrees*. The distracters are designed to reflect students' *most common misconceptions and misapplications of the concept being tested*.

Each of the distracters is given a *less-than-maximum point value depending on the proportion of correctness in the answer*.

One alternative is written that is completely wrong. This alternative is assigned *zero credit*.

All test item weights are linked to a Table of Specifications that justifies each test item alternative weight and each test item weight in relation to the *amount of time spent teaching the concept(s), and the number of concepts involved in each answer*.

---

**Weighted testing: Difficulty and discrimination index**

Test item statistics such as Difficulty and the Discrimination Index must be interpreted with caution in a weighted testing environment.

The variability in student responses (right or wrong) is radically altered in a multiple-correct-answer/weighted test.

For example, under a multiple-weighting scheme with a four-alternative multiple-choice test item, a student has a 75 percent probability of being correct as opposed to a 25 percent probability of being correct in a single-correct non-weighted (traditional) multiple-choice test item scheme.

This increased probability of correctness affects guess-correction calculations, and may also cause fluctuations in the variability of student responses to test items.

The fluctuation in the variability among student responses to items directly impacts Discrimination Index calculations.

A total test score in a weighted alternative multiple-choice test is the total item weight a student accumulates by answering certain questions correctly. A multiple-choice weighted alternative test will yield a larger score for a given student than if the test were non-weighted. This factor:

- Increases the possible range of student scores.

- Has the potential to increase the value of any test statistic which is computed on the basis of score variability.

---

---

**Weighted testing:  
Difficulty and  
discrimination  
index (Continued)**

The use of multiple-weighting schemes for tests will have effects on test item statistics and long-term stability characteristics of the test item statistics.

---

**Establishing  
passing scores**

*Establishing cut-off scores* for tests is directly related to test scoring methods. Guidelines for Standard-setting for Norm-referenced and Criterion-referenced tests are discussed in Section A of this Chapter.

Rigorous procedures exist for setting and validating cutoff scores for certification examinations, but not usually for other kinds of tests.

---

**Test and item  
analysis**

The issue of whether to rely on Norm-Referenced or Criterion-Referenced test and test item statistics was discussed in Section 2 of this Chapter. Criterion-referenced test and item analysis techniques have not been widely used, even though *most Air Force instructional systems are Criterion-Referenced*.

Some Criterion-Referenced test and item analysis statistics were difficult to generate because of the complex mathematics involved. Computer-based statistical analysis programs have mitigated this difficulty.

If Norm-Referenced statistics and Norm-Referenced interpretations are used to analyze Criterion-Referenced statistical data that are generated from a Criterion-Referenced instructional system, the interpretation of the test and item analysis data is usually erroneous.

---

**Erroneous  
applications of  
normative-  
referenced  
statistics on  
criterion-  
referenced  
applications**

This section describes some lessons learned from the erroneous application of Norm-Referenced statistics and interpretations to Criterion-Referenced test and test item statistics generated from a Criterion-Referenced instructional system.

Two of the more common examples of erroneous application and interpretation of statistics are the *use of the Discrimination Index for Criterion-Referenced test items*, and the *use of Reliability Coefficients for Criterion-Referenced tests*.

---

---

**Using the  
discrimination  
index for CR items**

Norm-Referenced (NR) interpretations of the Discrimination Index say that the higher the item discrimination, the better.

The function of a test in a NR instructional environment is to discriminate as much as possible among test takers in terms of relative intellectual (cognitive) and psychomotor (performance) skills.

The function of a test in a CR instructional environment is to discriminate as much as possible between those who master the instructional (course) objectives and those who do not.

---

**Differences  
between NR and  
CR discrimination  
index**

Using the same Discrimination Index to provide decision-making information in both NR and CR environments is problematic because the assumptions on which NR statistics are based do not apply in CR situations. Some of the differences between the NR and CR Discrimination Index are:

**NR Environments**

A NR Discrimination Index is typically computed by first rank ordering the entire tested population into roughly equivalent intervals (e.g., thirds or fifths).

Then, the number of students in the top interval getting an item correct is subtracted from the number of students in the bottom interval also getting the item correct.

Dividing this difference by an appropriate constant yields a Discrimination Coefficient ranging from -1 to +1.

This NR Discrimination Index is based on a normative distribution of raw (or percentage) scores, its interpretation must be based on that normative distribution.

**CR Environments**

In a CR instructional environment, the distribution of test scores is hardly ever statistically normal. The distribution of test scores is almost always negatively skewed (the bulk of the test scores are at the high end of the test score scale).

A negatively skewed CR test score distribution invalidates a NR interpretation that is based on a normal distribution.

---

---

**Differences  
between NR and  
CR discrimination  
index (Continued)**

A well performing CR instructional program will show most of the students performing with high scores on the test. This distribution of CR test scores will yield test items with relatively low (or no) discrimination if CR Discrimination Indexes are computed using NR formulas.

If a strict interpretation of the Discrimination Index using a NR interpretation is followed in a CR environment, test items with low Discrimination Coefficients (Discrimination Indexes) must be discarded.

Therefore, using NR statistics and interpretation of test item discrimination for a CR test will indicate the need to replace many CR test items that are probably functioning properly in the CR instructional environment.

**CR Indexes and Statistical Metrics**

Two Discrimination Indexes recommended for use with CR tests are Brennan's B Coefficient, and the Uninstructed-Instructed Group Difference Index.

Section A of this Chapter discusses the use of the Uninstructed-Instructed Group Difference Index, and the use of other CR-based statistics. CR statistical metrics are based on the proportions of masters and non-masters who answer test items correctly.

---

**Using reliability  
coefficients for CR  
tests**

*Use of Reliability Coefficients for Criterion-Referenced Tests*

An erroneous use of NR statistics is using a NR Reliability Coefficient for competency-based CR tests in competency-based courses of instruction.

The Kuder-Richardson formulas are widely used to determine test reliability.

The formulas are easily computed, and the coefficients can be readily interpreted.

*The formulas are based on a tenuous, and typically false assumption, that the underlying distribution of test scores is normal.*

The distribution of test scores is hardly ever normal for CR tests in a competency-based instructional system.

The distribution of test scores is not always normal even for NR tests in Norm-Referenced instructional systems.

---

---

**Using reliability coefficients for CR tests (Continued)**

Unless the testing situation involves a relatively large number of students (100 or more), the Kuder-Richardson assumption of a normal distribution may not hold true, even for NR tests in Norm-Referenced instructional systems.

A series of *reliability statistics* for CR test analysis (more correctly termed *dependability statistics*) have been developed that do not rely on assumptions about the shape or distribution of test scores.

---

**Distribution-free statistics**

These distribution-free statistics fall into two categories:

*Threshold-loss Indices*  
*Squared Error Loss Indices*

These statistical methods are discussed in Section B of this Chapter.

---

**Computer network test administration**

Administering tests via computer networks is a fairly recent development. Portions of Air Force and Navy education and training communities are experimenting with massed testing via computer networks.

This trend is extending to corporate and industrial training and education programs because of the availability of affordable network hardware and software.

---

**Recommendations for computer-based test administration**

*Lessons Learned for Test Administration Software*

Test administration software should have the ability to provide both diagnostic (formative) and evaluative (summative) tests.

*Diagnostic testing* requires examinees to be given detailed feedback on their incorrect responses.

*Evaluative testing* should not provide direct feedback to students other than a final test score.

*Diagnostic feedback* should also be available to examinees for off-line review.

---

---

**Recommendations  
for computer-  
based test  
administration  
(Continued)**

One way to do this is to provide the examinee with printed feedback information.

Another way is to provide the capability to save the feedback to a disk file that the examinee can review on the computer at a later time. This approach requires the capability to safeguard the information in order to ensure examinee confidentiality.

Another quality of test administration software is that it should post test results to a student record database without student intervention.

The software should give evaluation personnel the flexibility to access the test results, on an as required basis. Faculty can be permitted to access their student's test results without accessing the full student record database.

The software should accommodate the great variety of Interactive Courseware test questions (e.g., test questions that contain scenarios, animated or still graphics, text, digitized video, and digitized photographs and associated feedback).

---

**Additional  
information**

The following bibliography provides additional information on test item development:

- Berk, R.A. (1984). *A Guide to Criterion-Referenced Test Construction*. Baltimore, MD: Johns Hopkins University Press, Chapters 5 and 9.
- Brennan, R.L. (1972). A Generalized Upper-Lower Item Discrimination Index. *Educational and Psychological Measurement*, 32, 289-303.
- Brennan, R.L. and Kane, M.T. (1977). An Index of Dependability for Mastery Tests. *Journal of Educational Measurement*, 14, 277-289.
- Downing, S.M. (Fall 1992). True-False, Alternate-Choice, and Multiple-Choice Items. *Educational Measurement: Issues and Practice*, 11 (3), 27-30.
- Ebel, R.L. (1971). How to Write True-False Test Items. *Educational and Psychological Measurement*, 31 (2), 417-426.
- Ebel, R.L. (1979). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice Hall.
-

---

**Additional  
information  
(Continued)**

- Ebel, R.L. (1982). Proposed Solutions to Two Problems of Test Construction. *Journal of Educational Measurement*, 18 (4), 267-278.
- Ebel, R.L. and Frisbie, D.A. (1986). *Essentials of educational measurement*. Engelwood Cliffs, NJ: Prentice-Hall.
- Foltz, D. (1990). *NHCS Occasional Paper Number 3*, Washington, DC: National Home Study Council.
- Haladyna, T.M. (1991). Generic Questioning Strategies for Linking Teaching and Testing. *Educational Technology Research and Development*, 39 (1), 73-81.
- Haladyna, T.M. and Downing, S.M. (1989). Validity of a Taxonomy of Multiple-Choice Item-Writing Rules. *Applied Measurement in Education*, 2 (1), 51-78.
- Hambleton, R.K. and Novick, M.R. (1973) Toward an Integration of Theory and Methods for Criterion-Referenced Tests. *Journal of Educational Measurement*, 4, 106-126.
- Killoran, J. (February 1992). In Defense of the Multiple-Choice Question. *Social Education*, 56 (2), 106-106.
- Ornstein, A.C. (January-February 1992). *Essay tests: Use, Development, and Grading*, 65 (3), 175-177.
- Roid, G.H. and Haladyna, T.M. (1982). *A Technology for Test-Item Writing*. New York: Academic press.
-

## **Section E**

### **Summary of Lessons Learned for Test Item Development**

---

**Summary of  
lessons learned for  
test item  
development**

Table 12 (next pages) provides a summary of lessons learned for the various types of test items.

---

Table 12 Summary of Lessons Learned for Test Item Development

<b>Lessons Learned for All Test Items</b>
<p>Write the items in preliminary form during the instructional system development period.</p> <p>Use a test blueprint or outline to keep an appropriate relationship between the items on the test and the instructional objectives.</p> <p>Base each test item on an <i>important</i> point, idea, or skill.</p> <p>Write items to measure <i>understanding</i> or ability to <i>apply principles</i>.</p> <p>Test <i>one, and only one, point or idea per test item</i>.</p> <p>Write items that require <i>specific knowledge</i> of material studied, not items that require general knowledge or experience.</p> <p>Use clear and concise language that is <i>appropriate for the conceptual difficulty level</i> of the specific objective being tested.</p> <p>Present each test item task as simply and straightforwardly as possible.</p> <p>Keep test items free of extraneous, ambiguous, or confusing material.</p> <p>Keep test items free of tricky expressions, slang, or other tricky requirements.</p> <p>Review test items from other sources, such as textbooks, and other instructors.</p> <p>Use original language, not that found in textbooks or other instructional materials for the course.</p> <p>Eliminate any clues within the test item, or clues that relate to other items in the test.</p> <p>Be especially sensitive to clues or suggestions that could help a naive examinee (one who does not have the knowledge or skill that should be able to answer the item correctly).</p> <p>For tests that measure <i>discrimination, concrete concept, or defined concept intellectual skills</i>, make each test item independent of other items. Ensure that the answer to one test item is not dependent on the answer to other test items. (Some tests that measure <i>rule learning or problem solving intellectual skills, verbal information skills, cognitive strategy skills, or the memorization component of psychomotor skills</i> may be designed to require correct answers to a sequence of test items. These tests require that the correct answer to one, or a series of test items, is dependent on the correct answers to other previous test items).</p> <p>Ensure that test items are reviewed by other <i>instructors and content specialists</i> to help eliminate ambiguity, technical errors, or other errors in the test item.</p> <p>Ensure test items are reviewed by individuals who are <i>not content specialists</i> for ambiguity, clues for naive examinees, and for selected-response test items, plausibility for the naive examinee.</p> <p>Ensure that the test item has "face validity", measures a specific objective, and relates to the content studied in the course of instruction.</p> <p>Avoid the appearance of bias in the test item (e.g., race, gender, cultural, ethnic, regional, handicapped, age-group, or other apparent bias).</p> <p>Construct test items that have a clearly correct or clearly best answer.</p> <p>Follow standard rules of punctuation and grammar.</p> <p>For test items based on an opinion or authority, state whose opinion or what authority.</p> <p>Do not require <i>unnecessarily exact</i> or difficult operations. Test items should match <i>objective criterion standards</i>.</p> <p>Do not use specific determiners such as "always", "never", "none", and "all" in test items.</p> <p>Restrict the number of different item formats in a test. Use the most valid formats. Group items in the same format together.</p> <p>Use scenarios, pictorial material, or other graphics only when they are relevant to an objective or topic measured by the test item, and <i>only when required for the test item to effectively measure an intellectual or psychomotor skill</i>.</p> <p>When a scenario, picture, or graphic is used, provide specific test item directions referring to it.</p> <p>If scenarios are used for a test item, ensure that they are realistic and appropriate for the test item.</p> <p>If pictorial material or other graphics are used for a test item, ensure that they are clearly drawn and labeled.</p>

### Lessons Learned for Multiple-Choice Test Items

Write the stem so it *clearly defines the test item task*. Word the stem so that the examinee knows what is required *without seeing the response options*. Generally, writing the stem as a question helps to set the test item task more clearly.

When the stem is written as an incomplete statement, the option statements should complete the sentence, rather than beginning the item stem, or being inserted in the middle of the item stem.

Reduce the “reading load” as much as possible. Avoid repeating words in the option statements, by placing these words in the stem.

Do not provide verbal clues that point to the correct option or to elimination of incorrect option(s), such as disagreement between singular or plural, “a” and “an”, etc.

Make all option statements fit or match the stem.

*Have one, and only one, correct option.*

Make the options approximately equal in length. Avoid the tendency to make the correct option more detailed.

Make the options logically parallel, and about equal in complexity.

Make the options grammatically and syntactically parallel. Use grammatically and syntactically parallel words in the stem, the distracters, and the correct option.

Avoid using modifiers such as “sometimes” and “usually” in the options.

Ensure that each option has a unique meaning. Eliminate distracters with the same or similar meanings from the test item.

Make all distracters plausible to a naive examinee. Do not include implausible or impossible options as distracters in a test item.

Arrange the options in some appropriate, logical order.

Vary the position of the correct option.

*Avoid using “all of the above” as an option, and use “none of the above” sparingly.*

Avoid using negative words, including “except” in the stem and in the options. If it is necessary to use negative words, underline, capitalize, or highlight them for emphasis and examinee visibility.

Present the options in a vertical list. Each option should be on a separate line, beneath each other.

Use letters or numbers to label the options. Place the letters or numbers in front of the options

---

### Lessons Learned for Matching Test Items

Ensure that each of the sets is homogeneous in content. Ensure that the entries within a single set are logically and grammatically parallel.

Include more answer choices than statements that are to be matched with an answer choice.

Include only a reasonable number of answer choices for each test item (8-10).

Arrange the entries in the sets in some logical order.

Indicate whether an answer choice can be used more than once.

Provide directions that specify the basis on which the match is to be made.

Include separate headings for the statements and the answer choices.

Ensure that matching test items are the most appropriate format for measuring examinee mastery of the objective. Consider whether another format, such as multiple-choice test items, might be more appropriate.

Ensure that the matching test item is contained on a single page.

### Lessons Learned for True-False Test Items

Ensure that true-false test items are the most appropriate format for measuring examinee mastery of the objective. Consider whether another format, such as multiple-choice test items, might be more appropriate.

Ensure that the test item is *definitely true or definitely false*.

Ensure that the test item does not contain one part that is true, and another part that is false.

Ensure that the test item contains a single, important idea.

Keep the test item short.

Use simple language, if possible.

Ensure that there is not an insignificant word or phrase that influences the truth or falsity of the item.

*Do not use negative statements.*

Avoid the use of vague words such as “seldom” and “frequently”.

Do not use words that provide clues for the correct answer, such as “always”, “never”, “usually”, and “may”.

Ensure that true test item statements are not longer than false statements.

Balance the number of true and false test item statements in the test.

---

### Lessons Learned for Completion Test Items

Ensure that completion test items are the most appropriate format for measuring examinee mastery of the objective. Consider whether another format, such as multiple-choice test items, might be more appropriate.

Write the test item so that a single, brief answer is possible.

Keep the test item free of unimportant words.

Do not include so many blanks in the test item that the intent of the item is unclear.

Place the *blanks at the end* or near the end of the test item.

Avoid using specific determiners such as “a” and “an”, and singular or plural verbs in such a way that would clue the answer.

If the test item requires a numerical answer, indicate the units in which the answer is to be expressed.

If the test item requires a written answer, inform the examinee of the features which will be considered in scoring. For example, spelling, the number of words, or the format of the answer.

Write the test item clearly, so that there is only one possible correct answer.

Keep the length of the test item answer blank the same across all questions.

### Lessons Learned for Essay Test Items

Ensure that essay test items are the most appropriate format for measuring examinee mastery of the objective. Consider whether another format, such as multiple-choice test items, might be more appropriate.

*Use essay questions only to measure higher-order intellectual skills, such as rule-using, problem-solving, declarative (verbal) knowledge, cognitive strategies, and memorization components of psychomotor skills.*

*Use new material for transfer test items. Use previously presented material for production test items if the instructional objective requires memorization and written (or verbal) expression of the degree of internalization of the previously presented material.*

Match the test item question closely to the instructional objective it is supposed to measure.

Ensure that the test item question clearly defines the test item task for the examinee.

Allow ample time for answering the question. Inform the examinee of time restrictions if time is used as a metric for measuring examinee performance.

Inform the examinee of how answers will be scored and graded.

Try to start the test item question with a word or phrase such as “compare”, “contrast”, “give the reason for”, “give original examples of”, “explain how”, “predict what would happen if”, “criticize”, etc.

Provide a set of directions for the essays in the test.

If several essays are used in a test, include a range of complexity and difficulty in the essay test items.

*Prepare Criterion-Referenced standards for each essay test item before administering the test. Include key word, key phrase, time, and error metrics for each essay test item.*

---

**Rationale for establishing local norms for test statistics**

There is statistical and operational rationale for *establishing local norms for validation of test items* instead of relying on traditional Norm-Referenced indices for test item validation.

The conventional practice is to use the *Discrimination Index (DI) statistic to establish arbitrary norms* for validation of test items.

The Discrimination Index statistic is a measure of the effectiveness of a test item.

Discrimination indices range from a value of -1 to +1. Some conventional, arbitrary assumptions about the DI statistic for validation of test items are:

Any test item that displays a negative DI is assumed to indicate that students who do poorly on a test tend to get this item correct, while those students who do well on the test tend to miss the item.

Any test item that displays a negative DI is assumed to be flawed, and should be examined immediately. The test item must be revised or discarded, if necessary.

Any test item that has a large DI (closer to the value of +1) is more discriminating than a test item that has a smaller DI. The closer to +1 the DI of a test item is, the better the test item is. Test items with a DI farther away from +1 need to be revised.

---

**Local restrictions for discrimination index ranges**

There are several factors that can *restrict the range of the Discrimination Index in a local testing environment*:

The *nature of the curriculum* for which the test is written. Tests written at *different cognitive (intellectual) skill levels* will produce different test score distributions.

The *nature of the instructional environment*. Instructional environment factors that effect the range of the DI include:

The type of content.

The amount of practice.

The degree of feedback provided to the learners.

Rigor of the course.

The shape of the test score distribution, and the size of the sample tested will effect the statistical characteristics of all Norm-Referenced test and test item coefficients and indices.

---

---

**Diehl's procedure  
for establishing  
local norms for test  
statistics**

Dr. Grover Diehl investigated test item characteristics of Extension Course Institute (ECI) tests administered between 1974 and October, 1981. The following test item statistics were generated:

Average Discrimination Index (DI) values varied from a minimum of 0.06 to a maximum of 0.52.  
The average Grand Mean was 0.2618.  
The average Standard Deviation was 0.0487.

These average DI values fell within the "below average" range of ECI's existing DI norms for test items, which *assumed +1 as the maximum DI value*.

Using traditional Norm-Referenced indices for test item validation, it could be concluded that the ECI test items did not sufficiently discriminate, and therefore the test items should be revised or eliminated.

Using the concept of establishing local norms for test statistics, it could also be concluded that *local restrictions could require the development of local indices for test item validation*.

Based on local restrictions, Diehl recommended that ECI *revise its local DI range norms* as follows:

0 to 0.12	Undesirable
0.121 to 0.16	Low Differentiation
0.161 to 0.36	Average Differentiation
0.361 to 0.41	Above Average Differentiation
Above 0.41	High Differentiation

---

**Other uses of  
Diehl's procedure**

The procedure used to derive this new local norm range can be used by evaluation personnel to establish their own local test norms. The same procedure can be used (with minor modifications to the mathematics) to:

Determine the local ranges for Differentiation Indices.  
Determine the local ranges for Difficulty (Ease) Indices.  
Determine the local ranges for Reliability Coefficients

---

---

**Diehl's database**

Create a database of test scores and test item statistics. The larger the better.

The larger the database, the more stable computed ranges will be.

A larger database also tends to produce relatively normal distributions of data points.

---

**Diehl's distribution recommendations**

Gather data to describe the shape of the distribution:

**Normal Distribution**

If available, use a computer-based statistical package that can produce a histogram of data values to picture the shape of the data distribution. A relatively bell-shaped data distribution curve indicates a relatively normal distribution.

In a normal distribution, there is only one value that is most frequently occurring. This value is called the mode.

In a perfectly normal distribution, the mean, median, and mode are identical.

If computer-based statistical packages are not available, compare the arithmetic mean, median, and mode of the distribution. (Compare the Measures of Central Tendency.)

If the distribution of scores has only one mode, compare the mean and median values. If they are identical, the distribution is perfectly normal.

If the difference between the median and mode is a lot less than half the value of the mean, the distribution is closer to normal, and may be used with this procedure.

**Multi-modal Distributions**

If the data are multimodal (has more than one value occurring with equally high frequency), this procedure for establishing local norms for test statistics will not produce valid results. Calculate the range of the values by subtracting the lowest value from the highest value.

If the difference between the median and mode is more than half the value of the mean, the distribution is too skewed to be used with this procedure.

The closer the mean and mode values are, the better.

If the difference between the mean and mode values is larger than 1/4 of the mean value, look at the histogram of values and decide whether the distribution is close enough to normal to use.

---

---

**Diehl's procedure requires normal distribution of values**

The reason for insisting on a nearly normal distribution of values is because the procedure divides the distribution into segments based on the percentage of data points occurring in certain portions of the distribution.

The expected percentages of data points in certain portions of a standard normal distribution is well known.  
The simplicity of the procedure derives from the assumption of a normal distribution.

---

**Application of Diehl's procedure**

The first step in the procedure is to compute the *mean* and *standard deviation* (SD) of all the data points in the database.

For example, Diehl's study found a mean of 0.26, and a standard deviation of 0.05.

Discrimination indices based on correlation coefficients, such as the point biserial coefficient, cannot be simply averaged to obtain a mean value.

The *point biserial correlation coefficient* is a measure of the relation between a continuous variable such as scores on a test and a two-categorized, or dichotomous variable, such as "pass" or "fail" on a performance test item.

*The point biserial correlation coefficient is commonly used to measure the correlation between test scores and test items, or for measuring the correlation between test scores and psychomotor pass/fail performance criteria.*

To obtain a mean value for discrimination indices based on correlation coefficients, the following steps must be accomplished:

Transform the coefficient using a Fisher transformation.

Obtain the mean of the transformed values.

Retransform the resulting value into a correlation coefficient using a Fischer transformation.

Once the mean and standard deviation are computed, establish six cutpoints at 1, 2, and 3 Standard Deviations (SD) above and below the mean. *Section C, page 17, of this Handbook describes a grading scale with three cutpoints above and three cutpoints below a mean grade of 2.5.*

---

**Application of  
Diehl's procedure  
(Continued)**

On a normal distribution, this will account for more than 99% of all data points. This will effectively define the entire range of values for a particular test statistic. Using Diehl's mean value of 0.26 and standard deviation of 0.05 (rounded to two decimal places), the six SD cutpoints would be

-3 SD = 0.11	(0.26-0.15)
-2 SD = 0.16	(0.26-0.10)
-1 SD = 0.21	(0.26-0.05)
+1 SD = 0.31	(0.26+0.05)
+2 SD = 0.36	(0.26+0.10)
+3 SD = 0.41	(0.26+ 0.15)

Finally, the ranges need to be named to reflect the local norms for test statistics. Diehl used the following names and ranges for the *local discrimination index values for ECI test statistics*:

Undesirable	DI value less than -3 SD
Low	DI value between -3 and -2 SD
Average	DI value between -2 and +3 SD
High	DI value beyond +3 SD

**Lessons learned  
from local DI  
norms**

*Lessons Learned for Establishing Local Discrimination Index Norms*

The important point is that *local DI range norms that are established are based on local, empirical data, which is a function of local curricular conditions peculiar to a particular course of instruction.*

*The range of discrimination index (DI) values is highly restricted when compared to the theoretical limits of the DI (-1 to +1).*

In the example norms, a test item with a DI around 0.40 would be considered above average. Against the theoretical scale, it would be considered as a moderately discriminating item.

---

**Lessons learned  
from local DI  
norms (Continued)**

There is very little value in trying to improve a test item that is performing at the *upper end of its locally normal range*. The theoretical ranges found in textbooks are good starting points *when local empirical data are not available*.

If an instructional program is mature enough to have amassed a considerable base of test and test item performance data, seriously consider using the procedure for *establishing local, empirically-based norms for statistical coefficients and indices*.

Establishing local norms for tests and test item statistics will indicate if the local norms are sufficient to meet the objectives of a local instructional program, and decrease the reliance on arbitrary, or theoretical norms.

---

**Additional  
information**

The following bibliography provides additional information on establishing local norms for test statistics:

Diehl, G. (December 1981). *Characteristics of Kuder-Richardson Formula 21 Reliability Coefficients, Average Discrimination Indices, and Average Ease Indices Generated by ECI Tests* (January 1974 through October 1981). Air Force Extension Course Institute Research Report, Montgomery, AL.

Renckly, T.R. Toward establishing local norms for test statistics. *Distance Education and Training Council (DETC). News*, Spring 1994, pp. 19-22.

---

RICHARD E. BROWN III, Lt General, USAF  
DCS/Personnel

## Attachment 1 – GLOSSARY OF REFERENCES AND SUPPORTING INFORMATION

Walter Dick, Lou Carey, James O. Carey (2000). *The Systematic Design of Instruction 5<sup>th</sup> Edition*. Addison-Wesley Pub Co. ISBN 0321037804.

Patricia Smith & Tillman Ragan (1999). *Instructional Design, 2<sup>nd</sup> Edition*. John Wiley & Sons 399 pages. ISBN 047136570X.

Ruth Clark (1999). *Developing Technical Training 2<sup>nd</sup> Edition: A Structured Approach for Developing Classroom and Computer-based Instructional Materials*. International Society for Performance Improvement. 238 pages. ISBN 1-890289-C7-8

M. David Merrill (1994). *Instructional Design Theory*. Educational Technology Pub. ISBN 0-87778-275-X.

M. David Merrill, Robert D. Tennyson, and Larry O. Possey (1992). *Teaching Concepts: An Instructional Design Guide 2<sup>nd</sup> Edition*. Educational Technology Pub. ISBN 0-87778-247-4

Robert M. Gagné, Leslie J. Briggs, and Walter W. Wager. (1992). *Principles of Instructional Design 4<sup>th</sup> Edition*. Wadsworth Pub. Co. 365 pages. ISBN 00300347572

Robert M. Gagné (1985). *The Conditions of Learning and Theory of Instruction 4<sup>th</sup> Edition*. Holt, Rinehart and Winston. ISBN 0-03-063688-4

Jeroen J. G. van Merriënboer (1997). *Training Complex Cognitive Skills*. Educational Technology Publications.

Ruth Clark (1998). *Building Expertise: Cognitive Methods for Training and Performance Improvement*. International Society for Performance Improvement. 204 pages. ISBN 1890289043.

Bernice McCarthy (1996). *About Learning*. Excel, Inc. 452 pages. ISBN 0-9608992-9-4

Malcolm Fleming & W. Howard Levie (Editors) (1993). *Instructional Message Design: Principles from the Behavioral and Cognitive Sciences 2<sup>nd</sup> Edition*. Educational Technology Publications. 331 pages. ISBN 0-87778-253-9.

Ellen D. Gagné, Frank R. Yekovich, and Carol Walker Yekovich (1993). *The Cognitive Psychology of School Learning. 2<sup>nd</sup> Edition*. Addison Wesley Longman, Inc. 512 pages. ISBN 0673464164.

- Marcy P. Driscoll (1999). *Psychology of Learning for Instruction 2<sup>nd</sup> Edition*. Allyn & Bacon. 448 pages. ISBN 0205263216
- Ann E. Barron & Gary W. Orwig (1997). *New Technologies for Education: A Beginner's Guide 3<sup>rd</sup> Edition*. Libraries Unlimited. ISBN 1563084775
- Diana Laurillard (1993). *Rethinking University Teaching: A Framework for the Effective Use of Educational Technology*. Routledge. ISBN 0415092892.
- Tom Boyle and Tim Boyle (1996). *Design for Multimedia Learning*. Prentice Hall. 275 pages. ISBN 0132422158.
- William W. Lee and Diana L. Owens (2000). *Multimedia-Based Instructional Design: Computer-Based Training, Web-Based Training, and Distance Learning*. Jossey-Bass. 304 pages. ISBN 0787951595.
- Margaret Driscoll & Larry Alexander (Editor) (1998). *Web-Based Training: Using Technology to Design Adult Learning Experiences*. Jossey-Bass Inc. 288 pages. ISBN 0787942030.
- Brandon Hall (1997). *The Web-Based Training Cookbook*. John Wiley & Sons. 496 pages. ISBN 0471180211.
- Stephen M. Alessi & Stanley P. Trollip (2000). *Computer Based Instruction*. Allyn & Bacon. 432 pages. ISBN 0205276911.
- Andrew S. Gibbons & Peter G. Fairweather (1998). *Computer-Based Instruction*. Educational Technology. 570 pages. ISBN 0877783012.
- Douglas M. Towne (1995). *Learning and Instruction in Simulation Environments*. Educational Technology. 351 pages. ISBN 0877782784.
- Thomas M. Duffy & David H. Jonassen (Editors) (1992). *Constructivism and the Technology of Instruction: A Conversation*. Lawrence Erlbaum Associates. 221 pages. ISBN 0805812725.
- Brent G. Wilson (1995). *Constructivist Learning Environments: Case Studies in Instructional Design*. Educational Technology Publications. ISBN 0877782903.
- Roger C. Schank (Editor) (1997). *Inside Multi-Media Case Based Instruction*. Lawrence Erlbaum Associates. 451 pages. ISBN 080582538X.
- David H. Jonassen, Wallace H. Hannum & Martin Tessmer (1999). *Task Analysis Methods for Instructional Design*. Lawrence Erlbaum Associates. 275 pages. ISBN 0805830863.

- Allison Rossett (1999). *First Things Fast: A Handbook for Performance Analysis*. Jossey-Bass. 241 pages. ISBN 0787944386.
- Charles M. Reigeluth (Editor) (1983). *Instructional-Design Theories and Models: An Overview of their Current Status*. Lawrence Erlbaum Associates. 487 pages. ISBN 0-89859-275-5.
- Charles M. Reigeluth (Editor) (1999). *Instructional-Design Theories and Models: A New Paradigm of Instructional Theory. Vol. II*. Lawrence Erlbaum Associates. 715 pages. ISBN0-8058-2859-1.
- Tennyson, Robert D., Schot, Franz, Norbert, Seel & Dijkstra, Sanne. (Editors) (1997). *Instructional Design International Perspective Vol. 1 Theory, Research, and Models*. Lawrence Erlbaum Associates. 475 pages. ISBN 0-8058-1397-7.
- Sanne Dijkstra, Norbert Seel, Franz Schott & Robert D. Tennyson (Editors) (1997). *Instructional Design: International Perspective Vol. 2: Solving Instructional Design Problems*. Lawrence Erlbaum Associates. 418 pages. ISBN 0805814000.
- George M. Piskurich, Peter Beckschi, and Brandon Hall (Editors) (1999). *The ASTD Handbook of Training Design and Delivery: A Comprehensive Guide to Creating and Delivering Training Programs -- Instructor-led, Computer-based*. McGraw Hill. 530 pages. ISBN 0071353105.
- Charles R. Dills & A. J. Romiszowski (Editors) (1997). *Instructional Development Paradigms*. Educational Technology Publications. 882 pages. ISBN 08777882954.
- Sanne Dijkstra, Bernadette van Hout Wolters & Pieter C. van der Sijde (Editors) (1990). *Research on Instruction: Design and Effects*. Educational Technology Publications. ISBN 0877782210.
- David H. Jonnassen (Editor) (1996). *Handbook of Research on Educational Communications and Technology*. Macmillan. 1267 pages. ISBN 0028646630.
- Byron Reeves & Clifford Nass (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University Press. 305 pages. ISBN 1-57586-053-8.
- Donald A. Norman (1990). *The Design of Everyday Things*. Doubleday & Company. 256 pages. ISBN 0385267746,
- Bills, C.G., and Butterbrodt, V.L. (1992). *Total Training Systems Design Function: A Total Quality Management Application*. Wright-Patterson AFB, Ohio.

- Briggs, L.J., and Wager, W.W. (1981). *Handbook of Procedures for Design of Instruction* (2nd Ed.). Glenview, Illinois: Harper Collins Publishers
- Carlisle, K.E. (1986). *Analyzing Jobs and Tasks*. Englewood Cliffs, New Jersey: Educational Technology Publications.
- Davies, I.K. (1976). *Objectives in Curriculum Design*. London: Mc Graw Hill.
- Dick, W., and Carey, L. (1990). *The Systematic Design of Instruction* (3rd Ed.). Glenview, Illinois: Harper Collins Publishers.
- Gagné, R.M. (1985). *The Conditions of Learning* (4th Ed.). New York: Holt, Rinehart and Winston.
- Gagné, R.M., Briggs, L.J., and Wager, W.W. (1992). *Principles of Instruction* (4th Ed.). New York: Harcourt Brace Jovanovitch College Publishers.
- Gagné, R.M., and Merrill, M.D. (1990). *Integrative Goals for Instructional Design*. Englewood Cliffs, New Jersey: Educational Technology Publications. 38(1), 1-8.
- Goldstein, I.L. (1986). *Training In Organizations: Needs Assessment, Development, and Evaluation* (2nd Ed.). Pacific Grove, California. Brooks/Cole Publishing Company.
- Hageman, D.C. (1988). *Cognitive Engineering of Training Systems for Simulators*. National Aerospace and Electronics Conference. Dayton, Ohio.
- Hageman, D.C. (1985). *Effective Training Systems for High-Technology Equipment Operation*. National Security Industrial Association Fifth Annual Conference on Personnel and Training System Effectiveness. San Antonio, Texas.
- Keller, J.M. (1987). The Systematic Process of Motivational Design. *Performance and Instruction*, 26(9), 1-8.
- Kibler, R.J. (1981). *Objectives for Instruction*. Boston: Allyn and Bacon.
- Knirk, F.G., and Gustafson, K.L. (1986). *Instructional Technology: A Systematic Approach to Education*. New York: Holt, Rinehart, and Winston.
- Leshin, C.B., Pollock, J., and Riegeluth, C.M. (1992). *Instructional Design Strategies and Tactics*. Englewood Cliffs, New Jersey: Educational Technology Publications.
- Mager, R.F. (1962). *Preparing Objectives for Instruction* (2nd Ed.). Belmont, California: Fearon Publishers.

- Merrill, M.D., Tennyson, R.D., and Posey, L. (1992). *Instructional Design Strategies and Tactics*. Englewood Cliffs, New Jersey: Educational Technology Publications.
- Merrill, M.D., Lee, Z., and Jones, M.K. (1990). *Second Generation Instructional Design*. Englewood Cliffs, New Jersey: Educational Technology Publications.
- O'Neil, H.F., Jr., and Baker, E.L. (1991). Issues in Intelligent Computer-Assisted Instruction: Evaluation and Measurement." In T. Gutkin and S. Wise (Eds.), *The Computer and the Decision Making Process*. Hillsdale, New Jersey: Erlbaum Lawrence Associates.
- Reigeluth, C.M. (1983). Instructional Design; What is it and Why is it? In C.M. Reigeluth (Ed.), *Instructional Design Theories and Models? An Overview of Their Current Status*. Hillsdale, New Jersey: Erlbau Associates.
- Renckly, T.R. (April 1993) *Practitioner's Guide to Standard-setting*. Air University, Montgomery, AL.
- Renckly, T.R. (October 1990) *Criterion-Referenced Test Analysis: Another Look at a Compromised Process*. Paper presented at the 14th Annual Inter-Service Correspondence Exchange Conference, Pensacola, FL.
- Renckly, T.R. (April 1990) *A Guide to Criterion-Referenced Test Analysis*. Air University, Montgomery, AL.
- Renckly, T.R. (October 1989) *Test Validation: Current Practice--New Perspectives*. Paper presented at the 1989 Inter-Service Exchange Conference, Williamsburg, VA.
- Rossett, A. (1987). *Training Needs Assessment*. Englewood Cliffs, New Jersey: EducationalTechnology Publications.
- Spears, W.D. (1983). *Processes of Skill Performance: A Foundation for the Design and Use of Training Equipment*. (NAVTRAEQ-VIPCEN 78-C-0113-4). Orlando, Florida: Naval Training Equipment Center.
- Tennyson, R.D., and Michaels, M. (1991). *Foundations of Educational Technology; Past, Present and Future*. Englewood Cliffs, New Jersey: Educational Technology Publications.
- Wolfe, Petal. (1991). *Job Task Analysis: Guide to Good Practice*. Englewood Cliffs, New Jersey: Educational Technology Publications.

## Abbreviations and Acronyms

AFH	Air Force Handbook
AFM	Air Force Manual
CAI	Computer-Assisted Instruction
CBT	Computer-Based Instruction
CMI	Computer-Managed Instruction
CRT	Criterion-Referenced Test
CTS	Course Training Standard
ICW	Interactive Courseware
IG	Inspector General
JPR	Job Performance Requirements
OJT	On-the-Job Training
POI	Plan of Instruction
SME	Subject Matter Expert
STS	Specialty Training Standard
TCTO	Time Compliance Technical Order
TNA	Training Needs Assessment
TNR	Training Needs Assessment
TPT	Training Planning Team
TQR	Training Quality Report
TR	Training Requirement
U&TW	Utilization and Training Workshop

## Terms

The following list of definitions includes those terms commonly used to discuss education and training, including the test and measurement guidelines in this handbook. The list is not to be considered all-inclusive.

**Affective.** (see Attitude)

**Association.** The connection made between an input (stimulus) and an action (response).

**Attitude.** (a) The emotions or feelings that influence a learner's desire or choice to perform a particular task. (b) A positive alteration in personal and professional beliefs, values, and feelings that will enable the learner to use skills and knowledge to implement positive change in the work environment. Also see Knowledge and Skill.

**Behavior.** Any activity overt or covert, capable of being measured.

**Cognition.** The mental or intellectual activity or process of possessing intellectual skill knowledge, including associations, discriminations, discrete and concrete classifications, rule using and problem solving, verbal knowledge, and cognitive strategies.

**Cognitive Strategies.** (a) The capability of an individual to govern their own learning, remembering, and thinking behavior. (b) The ability of an individual to generate strategic and tactical behavioral decisions in response to judgments based on perceived and encoded cue (stimulus) conditions.

**Computer-Assisted Instruction (CAI).** The use of computers to aid in the delivery of instruction. A variety of instruction modes may be used, including tutorial, drill and practice, gaming, simulation, or combinations of instruction modes. CAI is an integral part of Computer Based Instruction (CBI), and Computer-Based Training (CBT).

**Computer-Based Instruction (CBI) and Computer-Based Training (CBT).** The use of computers to aid in the delivery and management of instruction. CBI and CBT are synonymous and are used interchangeably. CAI and CMI (Computer-Managed Instruction) are both elements of CBI and CBT.

**Computer-Managed Instruction (CMI).** The use of computers to manage the instructional process in CAI or CBT. Management normally includes functions such as registration, pre-testing, diagnostic testing and prescriptions of instructional materials, progress testing, and post-testing.

**Constraints.** Limiting or constraining conditions or factors, such as policy considerations, time limitations, equipment fidelity and availability, environmental factors, personnel, budgetary, or other instructional resource limitations.

**Course Chart.** A qualitative course control document that states the course identity, length, and security classification, lists major items of training equipment, and summarizes the subject matter covered in the course of instruction.

**Course Control Documents.** Specialized publications used to control the quality of the instructional system. Examples are Course Training Standards (CTS), Plans of Instruction (POI), Syllabus, and Course Charts.

**Courseware.** Education and training materials such as technical data, textual materials (such as lesson plans, instructor guides, student guides, and test and measurement instruments), audiovisual materials, and computer-based instructional materials.

**Criterion.** (a) The standard by which something is measured. (b) In the validation of test and measurement instruments, the standard against which test instruments are correlated to indicate the accuracy with which they predict human performance in some specified area. (c) In the validation and evaluation of instructional materials and systems, the measure used to determine the adequacy of a product, process, behavior, or other conditions.

**Criterion-referenced Test (CRT).** A test to determine, as objectively as possible, a student's achievement in relation to a criterion standard that is based on a criterion objective. During instructional system development, the CRT can be used to measure the effectiveness of the instructional system. The CRT may involve multiple-choice items, matching items, fill-in items, essay items, oral items, or performance items. If given immediately after the learning sequence, the CRT is an acquisition test. If given later in the learning sequence, the CRT is a retention test. If the CRT requires performance not specifically learned during instruction, the CRT is a transfer test.

**Diagnostic Test.** Instruments used to determine attainment of supporting skills and knowledge necessary to perform the terminal objectives. Diagnostic tests contain a number of test items in each specific subject area to allow a detailed search for a source of learning errors. Diagnostic tests are used during the validation (formative evaluation) of the instructional system to predict student success, and to identify and correct weaknesses in the instruction.

**Discrimination.** The process of perceiving and encoding a cue or stimulus and making a judgment concerning the condition of the cue or stimulus. A discrimination requires an individual to make judgments concerning the condition of a cue or stimulus and to respond differently to each condition.

**Duty.** (a) A large segment of work done by an individual. (b) Major divisions of work in a job.

**Enabling Objective.** Support or subordinate Criterion-Referenced objectives that must be accomplished by a learner in order to enable mastery of a terminal objective. See **Terminal Objective**.

**Evaluation.** A judgment expressed as a measure or ranking of instructional resources, including student achievement, instructor performance, the instructional process, the instructional system application, instructional materials, instructional equipment, training devices, and other factors. The types of evaluation include Formative Evaluation (Internal Review), Summative Evaluation (Operational Tryout) and Operational Evaluation (Internal Evaluation and External Evaluation).

**External Evaluation.** The acquisition and analysis of feedback data from outside the formal training environment to evaluate the graduate of the instructional system in an operational environment. Also called Field Evaluation. Also see Operational Evaluation.

**Feedback.** Information that results from or is contingent upon an action. The feedback does not necessarily indicate the correctness of an action; rather, it relates the results of an action from which inferences about the correctness of the action can be drawn. Feedback can be immediate, as when a fuel indicator instrument indicates the quantity of fuel in a fuel tank in real-time; or delayed, as when an instructor provides a discussion pertaining to an examination taken the previous week, or when completed graduate questionnaires are reviewed.

**Fidelity.** The degree to which an instructional system task, equipment, or training device represents the actual operational task, equipment, or device in terms of performance, characteristics, and environment.

**Field Evaluation.** See External Evaluation.

**Formative Evaluation (Internal Review).** An activity that provides information about the effectiveness of education and training materials for meeting education and training objectives and the acceptance of training materials as they are being developed. Formative evaluation includes individual, single-group, and small-group tryouts, and is the first step of the validation process for instructional materials and instructional systems. The purpose of formative evaluation is to make improvements to the instructional system or instructional materials while development is still in progress. Also called Developmental Testing. Also see Evaluation.

**Generalization.** Learning to respond to a new stimulus that is similar, but not identical, to one that was present during original learning. For example, during original learning, a child learns to call a beagle and a spaniel by the classification term “dog”. A child who can generalize would respond with the classification term “dog” when asked what kind of animal a schnauzer was.

**Instructional Objective.** See Objective.

**Instructional System.** An integrated combination of resources (students, instructors, materials, equipment, devices, and facilities), instructional techniques and procedures

that is capable of performing the functions required to achieve specified education or training learning objectives effectively and efficiently.

**Instructional System Developer.** A person who is knowledgeable of the instructional system development (ISD) process and is involved in the analysis design development, implementation, and evaluation of instructional systems. Also called Instructional Developer, Curriculum Developer, Education or Training Analyst, and other terms.

**Instructional System Development (ISD).** A deliberate and orderly, but flexible, process for planning, developing, implementing, evaluating, and managing instructional systems that ensures that personnel are taught the knowledge, skills, and attitudes required for job performance in a cost-effective manner. The ISD process depends upon a description and analysis of the tasks necessary for job performance, Criterion-Referenced objectives development, Criterion-Referenced test and measurement instruments developed before instruction begins, validation and evaluation procedures to determine if the objectives have been reached by the students, and methods for revising the instructional system or materials based on empirical data.

**Interactive Courseware (ICW).** Computer-controlled education or training designed to allow the student to interact with the learning environment through input devices such as mice, keyboards, joy sticks, or light pens. Student decisions and inputs to the computer determine the level, order, and pace of instructional delivery, and various forms of visual and aural instructional media.

**Job.** The duties, tasks and task elements (subtasks, steps and step actions) performed by an individual. The job is the basic unit used in carrying out the personnel actions of selection, training, education, classification, and assignment.

**Job Aid.** A checklist, procedural guide, decision table, worksheet, algorithm, or other device used by a job incumbent to aid in task performance. Job aids reduce the amount of information that personnel must recall or obtain in order to perform job tasks.

**Job Analysis.** The basic method used to obtain salient facts about a job, involving observation of workers, conversations with those who know the job, analysis questionnaires completed by job incumbents, or study of documents involved in performance of the job. Job analysis provides data on the missions, tasks, subtasks and steps required for job performance, the conditions and standards of performance, and the number of individuals and the time required for performing job components.

**Job Performance Requirements (JPR).** The missions, tasks subtasks, and steps required for the human component of job performance, the conditions under these job components may be performed, and the quality standards for acceptable performance. JPRs describe what people should do to perform their jobs.

**Knowledge.** Use of the cognitive processes (Intellectual Skills) that enable an individual to recall facts, make discriminations and judgments in response to perceived

cues and stimuli, identify discrete or concrete concepts, apply rules or principles, solve problems, state verbal information, or generate cognitive strategies or tactics in response to perceived or expected conditions. Knowledge is not directly observable. A person manifests knowledge through performing associated overt activities such as performance on a test or measurement instrument. Also see Attitude and Skill.

**Learning.** A change in behavior of the learner as a result of experience. The behavior can be physical or overt, or it can be intellectual or attitudinal.

**Lesson Plan.** An approved plan for instruction that provides specific definition and direction to the instructor on terminal and enabling learning objectives, equipment, training devices, instructional media requirements, and conduct of an education or training component of the instructional system. Lesson plans are a principal component of curriculum materials in that they sequence the presentation of learning experiences and program the use of supporting instructional materials, devices, and equipment.

**Media.** The delivery vehicles for presenting instructional material or basic communication stimuli to the student to induce learning. Examples of media include, instructors, textbooks, slides, audiovisual materials, video materials, interactive courseware (ICW) including ICW with multimedia capabilities, training equipment, and training devices.

**Metrics.** Measurement tools such as validation and evaluation instruments or test and measurement instruments used for assessing the qualitative and quantitative progress of instructional development with respect to the standards specified for instructional system development.

**Norm-Referenced Test.** The process of determining a student's achievement in relation to other students. Grading on the curve involves Norm-Referenced measurement, since a student's position on the curve (grade) depends on the performance of other students. Generally, Norm-Referenced measurement is not appropriate for the Air Force ISD process.

**Objective.** A statement that specifies precisely what learning behavior is to be exhibited by the student, the given conditions under which the behavior will be accomplished in the instructional system, and the minimum standards of learner performance required to demonstrate mastery of the objective. Objectives describe only the learning behaviors that directly lead to or specifically satisfy the skills, knowledge, and attitudes associated with a job performance requirement. An objective is a statement of instructional intent. Also referred to as Learning Objective or Instructional Objective. Also see Terminal Objective and Enabling Objective.

**Operational Evaluation.** The process of internal and external review of instructional system elements, instructional system requirements, instructional methods, courseware, training devices and equipment, facilities, and test and measurement instruments during full-scale operation of the instructional system. Operational evaluation is conducted to

determine if revision of the instructional system is required to enhance or to maintain system effectiveness and efficiency. Operational evaluation includes Internal Evaluation and External Evaluation. Also see Evaluation.

**Perceptual Skill.** The process of discriminating stimuli and associated cues encountered in the learning or job environment, and encoding the perceived cues or stimuli by making cognitive judgments as to whether a perceived condition is nominal, or whether it represents a condition that requires actions to respond to the condition. Also see Discrimination.

**Performance.** The part of a Criterion-Referenced objective that describes the observable student behavior (or the product of that behavior) that is acceptable to the instructor as proof that the student has accomplished a learning behavior to an acceptable standard of completeness or accuracy, and that learning has occurred.

**Plan of Instruction (POI).** A qualitative course control document designed for use within a school. The POI provides direction for course planning, organization, and operation. The POI contains Criterion-Referenced objectives, duration of instruction, required support resources and materials, and instructional guidance information for every instructional unit in an instructional system. The POI is also called a Syllabus.

**Post-test.** A Criterion-Referenced test designed to measure student performance on objectives taught during a unit of instruction. Used after exposure to an instructional program to provide a measure the changes that have occurred during instruction.

**Pre-test.** A Criterion-Referenced test designed to measure student performance on objectives to be taught during a unit of instruction and student performance on entry into the course of instruction. Used to measure the student's ability to attain each objective. A Readiness Pre-test used to measure prerequisite course entry skills. A Placement Pre-test (Adaptive Pre-test) used to measure attainment of course or unit objectives. Can also be used after the instructional system becomes operational to determine how much instruction individual student's need. A Diagnostic Pre-test is used to determine attainment of supporting skills and knowledge necessary to perform the terminal objective. Can also be used during validation to predict success, and to identify and correct weaknesses in the instruction. Diagnostic pre-tests contain a number of test items in each specific subject area to allow a detailed search for a source of learning errors. A Survey Pre-test is used to determine what prospective students already know and can do before receiving instruction. Survey pre-tests are used during development of instruction to gather data for design of instruction.

**Reliability.** (a) A characteristic of evaluation which requires that testing instruments yield consistent results. (b) The degree to which a test and measurement instrument can be expected to yield the same result upon repeated administration to the same population. (c) The capability of a training device, equipment or instructional system to operate effectively for a period of time without a failure or breakdown.

**Skill.** The ability to perform a job-related activity that contributes to the effective performance of a task. Skills involve physical or manipulative activities that require intellectual skills (knowledge) for their execution, and that also have specific requirements for speed, accuracy, or coordination. Also see Attitude and Knowledge.

**Subject Matter Expert (SME).** (a) An individual who has thorough knowledge of a job, including the job duties and tasks and the associated intellectual skills. (b) An individual who has thorough knowledge of a particular topic. (c) An individual who has thorough knowledge of a job or topic that qualifies the SME to assist in the instructional development process (for example to consult, review, analyze, advise or critique). (d) A person who has high-level knowledge and skill in the performance of a job.

**Summative Evaluation.** The operational tryout of an instructional system at the completion of the development process. Summative evaluation is conducted after the instructional system has become operational. Summative evaluation is the final step in the validation process. Data gathered during summative evaluation is used to determine the effectiveness of the instructional system, and identifies how well graduates can meet specified job performance requirements.

**Syllabus.** (See Plan of Instruction)

**System Approach to Training (SAT).** Procedures used by instructional developers to develop instruction. Each SAT phase requires input from the prior phase and provides input to the next phase. Evaluation provides feedback that is used to revise instruction. Also see Instructional System Development (ISD).

**Target Audience.** (a) The total collection of possible users of a given instructional system. (b) The persons for whom the instructional system is designed.

**Task.** A unit of work activity or operation that forms a significant part of a duty. A task usually has clear beginning and ending points and directly observable or otherwise measurable processes. Tasks frequently, but not always, result in a product that can be evaluated for quantity, quality, accuracy, or fitness in the work environment. A task is performed for its own sake, and is not dependent upon other tasks. Task performance may be sequential with other tasks that make up a duty or job array.

**Task Analysis.** The process of describing job tasks in terms of Job Performance Requirements (JPR), and the process of analyzing the JPRs to determine Training Requirements (TR). Also see Job Performance Requirements (JPR).

**Terminal Objective.** An objective the learners are expected to accomplish upon completion of the instruction. Terminal Objectives are composed of Enabling (support or subordinate) Objectives. Also see Enabling Objective.

**Test Validity.** The degree to which a criterion test actually measures what it is intended to measure.

**Training.** A set of events or activities presented in a structured or planned manner, through one or more instructional media, for the attainment, retention, and transfer of skills, knowledge, and attitudes required to meet Job Performance Requirements (JPR).

**Training Needs Assessment (TNA).** The study of operational job performance, and the job environment that influences job performance in order to make recommendations and decisions on requirements for training to close the gap between the desired job performance and actual job performance.

**Training Planning Team (TPT).** An action group composed of representatives from all pertinent functional areas, disciplines, and interests involved in the life cycle design, development, acquisition, support, modification, funding, and management of a specific defense instructional system.

**Training Requirements (TR).** (see Task Analysis)

**Training Strategy.** An overall plan of activities to achieve an instructional goal.

**Training System.** A systematically developed curriculum including, but not limited to, an integrated combination of resources (students, instructors, materials, equipment, devices, and facilities, and personnel to operate, maintain or employ the system), and instructional techniques and procedures that is capable of performing the functions required to achieve specified training learning objectives effectively and efficiently. A training system includes all necessary elements of logistic support. Also see **Instructional System.**

**Training (Instructional) Validity.** A metric for an instructional system that measures the degree to which students learn during exposure to the instructional system.

**Transfer Validity.** A metric for an instructional system that measures the degree to which what has been learned in the instructional system transfers as enhanced performance to the job environment.

**Utilization and Training Workshop (UT&W).** A forum to determine Specialty Training Standard (STS) requirements and responsibilities for a specialty. Workshop attendees include, but are not limited to, representatives from the training and using organizations.

**Validation.** The process of developmental testing, field testing, and revision of the instruction to be certain the instructional intent is achieved. The instructional system is developed unit by unit and tested (or validated) on the basis of the objectives for each instructional unit. Validation activities include Formative Evaluation (Internal Review), Summative Evaluation (Operational Tryout), and Operational Evaluation (Internal Evaluation and External Evaluation).