
# A Guide for Writing and Improving Achievement Tests

Teresa L. Flateby, Ph.D.

University of South Florida
Tampa, FL 33620
(813) 974-2742

# INTRODUCTION

Achievement tests can be written to ascertain students' level of learning within a course, in a major, or across their entire undergraduate education. For test results to be useful, they must follow basic measurement standards. In this document, procedures are presented which should help produce a valid test reflecting appropriate coverage of content and a reliable test with repeatable results. Both qualities are important for the two most widely used types of items, essay and multiple-choice, which are compared. Examples of printouts available from the Office of Evaluation and Testing are described and used to explain the item and test evaluation processes in the multiple-choice examination discussion.

Although a similar process is followed when designing a test for a single course or program, multiple faculty and a measurement consultant should be involved when designing a test to measure achievement within a major.

## Fundamentals of Achievement Testing

*The purposes of classroom achievement tests and their results are many and varied.*

Some of the possibilities are to:

- measure an individual's achievement of course objectives
- assess the group's performance
- evaluate the test and the items
- evaluate and improve instruction and the curriculum

**Always remember, however, that the fundamental purpose of achievement testing is to promote learning.**

Achievement test results should accurately measure individual differences or achievement at a certain pre-specified mastery level and should always foster learning.

To accomplish these purposes, a test must be valid and reliable. Validity is addressed when a test plan is formulated to accurately represent the course content and depth of learning achieved in a course. Test results must be reliable or repeatable to be confident that a student's score is a true reflection of an examinee's achievement. When a test is constructed which closely adheres to the test plan and other guidelines presented in this manual, the likelihood of gaining repeatable test results that accurately reflect achievement of the course content is improved.

Several testing factors have been shown to contribute to learning. First, the type of test students expect guides study behaviors. If a multiple-choice test is planned, students typically will study only for recognition. If it is known that a test will emphasize factual information, a student will memorize facts, which usually are forgotten quickly. Second, test questions written above the rote level have a greater potential for promoting transfer and retention. Therefore, most tests should be written to include items to stimulate higher cognitive levels.

**Essay or Multiple-Choice Achievement Test?**

There have been many criticisms directed toward multiple-choice tests. It is often heard that they only measure rote learning. Moreover, they are often referred to as "multiple guess" exams. Thus, essay examinations are deemed necessary to measure cognitive levels above the simplest level of learning. However, measurement experts believe that these criticisms are unwarranted if a multiple-choice exam is well constructed.

Because the assessment of student learning requires an adequate and accurate sampling of course content, the multiple-choice test is recommended for achievement-

type testing. If the achievement test requires measuring the highest cognitive by combining both essay and multiple-choice items into a single test is commonly suggested. This is advocated especially if the test writer is new to achievement test construction.

Table 1 illustrates the appropriate uses of multiple-choice and essay examinations and the strengths and weaknesses of each.

## TABLE 1

*A Comparison of Essay and Multiple Choice Tests*

| | **Essay** | **Multiple Choice** |
|---|---|---|
| **Recommended Uses** | 1. When measuring the highest cognitive levels (synthesis and evaluation levels of Bloom's Taxonomy – Appendix A).<br>2. When a response needs to be created.<br>3. When evaluating writing ability. | 1. When measuring achievement at the knowledge, comprehension, application and analysis cognitive levels. |
| **Advantages** | 1. Relatively short amount of time required to construct the items.<br>2. Allows for creativity, originality and composition. | 1. Objective scoring. (Once "correct" answers are decided).<br>2. Evaluation of validity is possible by comparing the test to the table of item specifications.<br>3. Evaluation of reliability is possible.<br>4. Thorough sampling of course content is possible.<br>5. Item analysis resulting from test scores can reveal particular problems in the exam and/or in the instruction or learning. |
| **Disadvantages** | 1. Objective scoring is questionable, and more difficult.<br>2. No generally acceptable criteria for demonstrating the validity and reliability of test.<br>3. Course sampling is limited.<br>4. Time consuming to evaluate responses. | 1. Time consuming to construct.<br>2. Difficult to construct items at the highest cognitive levels.<br>3. Faculty must have some training or knowledge in test construction and item analysis techniques to write valid and reliable test. |

# ESSAY EXAMINATIONS

## Developing Essay Items

These essay development suggestions should maximize the effectiveness of essay items for measuring achievement of course content. For guidance on assessing the quality of writing, refer to *Cognitive Level of Quality Writing Assessment: Building Better Thought Through Better Writing, 2001*.

1. Use a table of item specifications, also called a test blueprint (discussed in "How to Construct a Valid Multiple-Choice Test" section), to ensure that items are relevant and appropriate for the course content.

2. Prepare students for taking an essay exam. Provide practice in writing essay responses. Score these and give feedback to the students about their responses. Give students the grading criteria before the test.

3. Focus the questions. Be precise so that students clearly understand what is expected of them.

4. Have all students respond to the same essay questions; do not let them choose among the questions. Course content cannot be adequately sampled if students select content on which they wish to be tested. Also, students' performance cannot be compared if they are tested on different content.

5. Write more essay questions that allow for restricted responses rather than one or two essay questions that require long responses. This improves content sampling, which is especially important if the essay is being used to measure achievement rather than writing ability. Validity and reliability are improved if content is accurately and adequately sampled.

6. Students should have sufficient time to plan, prepare, and review their responses. Consider this when planning the number of essay items to include on the test.

7. Have a colleague review the questions for ambiguities.

8. Write items that measure the application, analysis, synthesis and evaluation levels of Bloom's Taxonomy of Educational Objectives. (See Appendix A.)

## Scoring Essay Items

Although essay exams have been criticized for being unreliable, procedures exist which, when followed, can improve the consistency of scoring and therefore their reliability.

1. Review lecture notes and course materials before scoring students' essay responses.

2. Read each individual's response to a single item one time before scoring and before reading responses to the next item.

3. Have students sign their names on the backs of the papers so the examinees are anonymous.

4. Know what should be contained in each response before reading any papers. Specify the content to be covered. Also, determine the weight to be given to each element expected. Allow for unanticipated, but valid, responses. This is the reason for reading students' responses to an item one time before actually scoring them.

5. If achievement of the content is the sole emphasis of the course, ensure that achievement and not writing ability is being evaluated. Many measurement experts believe that sentence structure, grammar and other aspects of writing should not be considered in the scoring of a paper unless they are part of the course content. The general measurement perspective is that students should be tested only on the material taught in the course. Other educators disagree. If writing will factor into a student's grade, the importance of writing skills should be clearly emphasized before students prepare for the test.

In short, when developing essay questions for achievement exams, course content should be adequately sampled and expected responses should be specified as precisely as possible. Therefore, for testing achievement, more questions, restricted responses and specified response criteria are recommended.

# MULTIPLE-CHOICE EXAMINATIONS

## How to Construct Valid and Reliable Classroom Achievement Tests

Certain guidelines should be followed when developing an exam to adequately and accurately measure achievement of course content and to achieve reliable scores. Constructing a valid test which reflects course objectives is an integral part of planning the course. Thus, "teaching to the test" is desirable, even necessary. Also, confidence in test scores is imperative for assessing differences in learning, assigning grades, or for determining mastery. Instead, a test should produce repeatable scores. A 60 earned by a student will remain about 60 if a reliable test is repeated or an equivalent form of the test is given. In other words, a 60 is a close approximation of the student's "true" or theoretical score.

Achievement tests are either norm-referenced or criterion-referenced. Norm-referenced tests emphasize individual differences, how students compare with each other; criterion-referenced tests highlight how examinees' performance compares to a specific standard or level of mastery, logically or empirically determined. Identification of this standard is sometimes difficult to accomplish, especially at the more complex learning levels. Although difference in orientation exists between norm and criterion-referenced tests, Hopkins, Stanley and Hopkins (1990) maintain that all good achievement tests should be based on either explicit or implicit objectives or topics reflected in a table of item specifications. This implies that there is a great deal of overlap between the two types of tests and the development of each type begins similarly. The differences pertain mostly to the presentation and interpretation of the

results. These similarities and differences will be discussed further in the Item Analysis section.

A well-constructed test blueprint, also referred to as a table of item specifications, provides the necessary structure to foster validity. (Tables 2-5 are examples of test blueprints.) Thus, to measure achievement of a unit's, course's, or program's objectives, the test blueprint must be an accurate representation of the content and cognitive levels taught.

## TABLE 2

| | TASK | | | |
|---|---|---|---|---|
| **CONTENT** | **Knows Specific Facts** | **Understands Concepts** | **Applies Principles** | **TOTALS** |
| Newton's Laws of Motion | 4 | 4 | 12 | **20** |
| Types of Forces | 4 | 2 | 7 | **13** |
| Buoyancy | 2 | 4 | 4 | **10** |
| Acceleration of Gravity | 2 | 3 | 5 | **10** |
| Friction | 2 | 2 | 3 | **7** |
| **TOTALS** | **14** | **15** | **31** | **60** |

*Zimmerman, B.B., Sudweeks, R.R., Shelley, M.F., Wood, Bud, 1990.*

## TABLE 3

| | KNOWS | | | | | |
|---|---|---|---|---|---|---|
| **OUTCOMES CONTENT** | **Terms** | **Facts** | **Procedures** | **Comprehends Principles** | **Applies Principles** | **Total Number of Items** |
| Role of Tests in Instruction | 4 | 4 | | 2 | | **10** |
| Principles of Testing | 4 | 3 | 2 | 6 | 5 | **20** |
| Norm-Referenced versus Criterion-Referenced | 4 | 3 | 3 | | | **10** |
| Planning the Test | 3 | 5 | 5 | 2 | 5 | **20** |
| **Total Number of Items** | **15** | **15** | **10** | **10** | **10** | **60** |

*Gronlund, 1982.*

## TABLE 4

| | TAXONOMY LEVEL | | | |
|---|---|---|---|---|
| **MAJOR CONTENT STRATA** | **Knowledge** | **Comprehension** | **Application, Synthesis, etc.** | **TOTAL** |
| The functions of measurement in education | 3 | 2 | 0 | **5 (10%)** |
| Basic statistical concepts, central tendency and variability | 1 | 2 | 2 | **5 (10%)** |
| Norms: types, meaning, interpretation | 3 | 3 | 4 | **10 (20%)** |
| Validity: content, construction, criterion-related validity and correlation | 4 | 6 | 5 | **15 (30%)** |
| Reliability: concepts, theory, and methods of estimation | 4 | 7 | 4 | **15 (30%)** |
| **TOTALS** | **15 (30%)** | **20 (40%)** | **15 (30%)** | **50 (100%)** |

*Hopkins, Stanley, Hopkins 1990.*

## TABLE 5

| | LEARNING OUTCOMES | | | |
|---|---|---|---|---|
| **CONTENT** | **Knowledge** | **Comprehension** | **Application and above** | **Total Number of Items** |
| **Purposes of Testing** | 3 | 2 | | **5 (10%)** |
| **Necessary Criteria for Tests** | | | | |
| Reliability | 2 | 1 | | **3 (6%)** |
| Validity | 2 | 2 | | **4 (8%)** |
| **Test Development** | | | | |
| Table of Item Specifications | 2 | 3 | 3 | **8 (16%)** |
| Proper Item Construction | 5 | 3 | 2 | **10 (20%)** |
| **Criteria for Evaluating Test** Relevance, Variability, Difficulty, Discrimination, Reliability | 3 | 2 | | **5 (10%)** |
| **Item Analysis** | | | | |
| Principles/Printout | 3 | 2 | 2 | **7 (14%)** |
| Discrimination, Difficulty, Distractor Analysis | 4 | 3 | 1 | **8 (16%)** |
| **TOTALS** | **24 (48%)** | **18 (36%)** | **8 (16%)** | **50 (100%)** |

*Test Construction Workshop, Flateby.*

To construct a test blueprint, first list the important course content, which are reflected in the syllabus and lesson plans. These will be listed on the far left column. Next, determine the cognitive levels of understanding students should achieve for each of the content areas. Bloom's Taxonomy of Educational Objectives and Cognitive Domain (1956), or a similar hierarchy, is typically used to specify the depth of learning expected. How thoroughly should students understand the material? Should they be able to recognize an appropriate step in a process (knowledge), explain a concept in their own words (comprehension), apply a principle or process to a new set of circumstances (application), compare and contrast components of schema (analysis), or create a plan to solve a problem (synthesis)? Knowledge, the foundation of Bloom's hierarchy, represents remembering or recognizing facts, followed by comprehension, the basic level of understanding. At the application level, a learner uses the content, skill or concept learned in a situation not encountered in class, the readings, or assignments. Analysis, the fourth level, requires a person to divide the material, a concept or process, into its component parts, to interrelate, compare, and contrast the parts. The fifth level, synthesis, involves the combination of components into a whole product, plan, or procedure. In the highest cognitive level, evaluation, a person judges a product or process based upon a specific set of criteria. (See Appendix A for a complete description of Bloom's Taxonomy.) The cognitive levels provide headings for the next columns. Typically, the highest cognitive levels are grouped into a single heading. (Refer to Tables 2–5.)

After the content and cognitive levels have been specified and placed on the table, determine the percentage of items to be assigned to each of the content areas

and cognitive levels. These percentages are based upon the importance of the content, the emphasis given the content in the course or program, the potential it has to increase the retention and transfer of learning, and the cognitive levels fostered in classroom assignments. Calculate the number of items to accompany the percentages based upon the total number of test items. When deciding upon the total number of test items, keep in mind that all students should have adequate time to finish the exam, but reliability is usually strengthened with well-written items. Also, allow at least one minute for each item written above the knowledge level.

Consideration of several other factors should help produce valid test results. There should be no surprises on the test. If only facts were presented in class and in the assignments, do not include analysis-type questions. Similarly, if concepts were analyzed, write an appropriate number of items requiring analysis, which could be short essay-type items. If the test blueprint is reflective of the content and cognitive levels, and is followed carefully when writing the items, a reliable assessment of students' achievement should result.

Table 5 represents a plan to measure learning from a one-day test construction and item analysis workshop. The important content topics are listed in the first column and the cognitive levels are presented in the next three columns. The Item Analysis content area (discrimination etc.), items will be written to elicit multiple cognitive levels, four items at the knowledge or factual level, three items at the basic understanding level and one item at the application level. These numbers suggest that students were given little time to apply the information presented.

To summarize the steps for constructing a test blueprint to achieve a valid achievement test:

1. List important course content or topics.

2. Identify appropriate cognitive levels using Bloom's Taxonomy of Educational Objectives for each of the course objectives.

3. Determine the number of items for
    a. the entire test
    b. each cell, i.e. course content by cognitive level.

**Addressing Reliability**

Reliability coefficients, appropriate for norm-referenced exams, are typically calculated by correlating or comparing two sets of scores. The most appropriate method to estimate reliability of a classroom achievement test is the Kuder-Richardson 20 (KR-20), an internal consistency measure which relates scores within one administration. Basically, the KR-20 is calculated by comparing the totals of the correct and incorrect responses for each item (the sum of the individual item variances) to the total test variance.

A reliability coefficient can range from 0.0, representing no consistency, to 1.00, representing perfect consistency. Typically, a reliability coefficient of at least .70 or higher is considered necessary to place confidence in the scores of a norm-referenced achievement test, which is critical for assigning grades. Below .70 it is less probable that scores are attributed to achievement rather than to chance or testing error. By adhering to the following guidelines, the likelihood of constructing a reliable norm-referenced test with results reflecting a normal distribution is increased.

1. Write longer tests, with well-constructed items. (Refer to Developing Items.)

2. Include items which are positive discriminators. This means that, in general, students who perform well on the exam answer the item correctly.

3. Write items which are moderately difficult. Because variation in scores contributes to reliability, very easy and very difficult items do not add to reliability as much as items of moderate difficulty and also have less potential to discriminate. However, a few easy items at the beginning of the test might build examinees' confidence. (Refer to Evaluating Tests and Items.)

Reliability coefficients calculated by correlating two sets of scores will be lower for a minimum competency criterion-referenced test than for a norm-referenced test because more students should answer items correctly, resulting in less variation. Also, these results should form a negatively skewed distribution, with most scores clustering at the high end of the distribution. Therefore, a different set of criteria is necessary to evaluate the consistency of scores for criterion-referenced exams. Hopkins, Stanley, and Hopkins (1990) recommend using the standard error of measurement as an indicator of score consistency, which is discussed in the Item Analysis section.

To summarize, the appropriate statistical method to estimate reliability and the acceptable level will vary depending upon the intent of the achievement test. If the test was developed strictly to measure individual differences, the reliability coefficient should be .70 or above. Although variation, and ultimately reliability is enhanced by including moderately difficult terms, it is acceptable to begin with a few easier items to promote confidence and to end with more difficult items to challenge the better prepared students. If a mastery-level or a minimum competency criterion-referenced test was constructed, less variability is expected because more students should answer the items correctly. Thus, a KR-20 calculated for a criterion-referenced test is expected to be weaker if mastery is achieved.

## Developing Items

Before reading this section, take a moment to answer the questions in Activity 1.

Compare your answers with the correct answers provided in Appendix B.

*Activity 1: Testing Your Multiple-Choice Test-Wiseness*
*(Adapted from Eison, 1985)*

DIRECTIONS: The seven multiple-choice questions below cover historical topics that you are not likely to know. See if you are able to determine the correct answer by carefully reading each item. Please circle the correct answer for each item.

1. The Locarno pact:
    a. is an international agreement for the maintenance of peace through the guarantee of national boundaries of France, Germany, Italy, and Belgium.
    b. allowed France to occupy the Ruhr Valley.
    c. provided for the dismemberment of Austria-Hungary.
    d. provided for the protection of Red Cross bases during war times.

2. The disputed Hayes-Tilden election of 1876 was settled by an:
    a. resolution of the House of Representatives.
    b. decision of the United States Supreme Court.
    c. Electoral Commission
    d. joint resolution of Congress.

3. The august character of the work of Pericles in Athens frequently causes his work to be likened to that in Rome of:
    a. Augustus.
    b. Sulla.
    c. Pompey.
    d. Claudius.

4. The Declaration of the Rights of Man was:
    a. adopted by the French National Assembly.
    b. adopted by every Western European legislature.
    c. immediately ratified by every nation in the world.
    d. hailed by every person in England.

5. The Locarno pact:
    a. was an agreement between Greece and Turkey.
    b. gave the Tyrol to Italy.
    c. was a conspiracy to blow up the League of Nations' building at Locarno.
    d. guaranteed the boundary arrangements in Western Europe.

6. Horace in the 16th Epode emphasizes the:
    a. despair of the average man confronted by sweeping social change.
    b. elation of the average man confronted by sweeping social change.
    c. optimism of the common man about sweeping social change.
    d. all of the above.

7. About what fraction of the 1920 population of the United States was foreign-born?
    a. less than five percent.
    b. between fourteen and twenty-eight percent.
    c. twenty-five percent.
    d. between thirty and fifty percent.

Some of these items may have been answered correctly even with little or no exposure to the content. All of these "clues" should be avoided when developing items. Many "test-wise" students are able to guess the correct answers to all of the items in the exercise and would make the same guesses if the same items were administered a second time. Thus, the test would be reliable, but the results would not be valid as a measure of achievement of course content. By following the test blueprint and the guidelines offered on the next pages, the chances of achieving valid and reliable test results are increased.

**General Item Writing Guidelines**

An item or question contains three parts:

- the stem, in which the question is asked or the problem is stated
- the correct option
- the incorrect options, also called foils or distractors.

The item should have only one correct answer and should be based upon significant information or concepts, not trivia. The item also should be clearly defined and be worded precisely without ambiguities. Remember, reading achievement is not being tested unless that is what is being taught, so be as brief and concise as possible.

To achieve a test with valid results, adhere to the test blueprint and ensure the items are written to reflect the appropriate cognitive levels. A test which promotes retention and transfer has items written at various levels in Bloom's Taxonomy (or some other defensible learning taxonomy). Be certain that items are written to accurately reflect the cognitive levels encouraged in the course for each important content area. Appendix C presents verbs which are appropriate for the various cognitive levels in Bloom's Taxonomy and may be useful when developing items. Follow the guidelines below when developing multiple-choice test items.

*Write the stem:*

1. as a complete sentence or question, or an incomplete statement which is completed by selecting one of the responses. It is easier to write complete statements or questions without ambiguity. Measurement experts recommend complete sentences for those new to test construction.

2. in a positive form. Negative items are easier to write and easier for students to answer. For example, "Which of the following does not promote reliability in a norm-referenced test?"

3. with a single correct answer. The stem may ask for the best answer, which elicits finer discriminations.

4. as precisely as possible. However, given a choice between a longer stem or longer options, lengthen the stem.

5. in more detail, instead of lengthening the options. When words are repeated in the options, lengthen the stem to include the repeated words. Attempt to write the stem as briefly as possible.

*All options, both incorrect (called distractors) and correct, should:*

- be brief.
- be grammatically consistent with the stem.
- be approximately the same length.
- be equally complex.
- cover the same type of content.
- be independent of each other.
- follow the rules of grammar.

*The distractors should:*

- be written for students who have a partial understanding or misunderstanding of the content.
- be plausible.
- be similar to the correct answer.

It is unnecessary to have the same number of distractors in all of the items. Stop adding distractors when they are no longer plausible. The inclusion of implausible distractors only increases the amount of reading time and does not add anything to the item or test.

*The correct responses should not:*

- provide clues, be longer, more technical or repeat any important words from the stem.

It is typically advised to refrain form using “all of the above” because a student is able to select the correct response from partial information. If he or she knows two answers are correct, “all of the above” will be selected even if he or she is unsure of the other options. Also, if a student knows one of the options is incorrect, then “all of the above” will not be selected. “None of the above” merely shows that a student is able to identify what is incorrect but does not provide evidence that the accurate information is known.

**Evaluating Tests and Items**

The following criteria adapted from Ebel and Frisbie (1991) provide a useful framework for evaluating norm-referenced achievement tests and items. If the evaluation processes or criteria are different for criterion-referenced tests, a separate description is presented.

Relevance: Does an item belong in the test? Do the items measure what the test author intended to measure? These are validity questions and address the extent to which the test blueprint was used.

Balance: Do the items adequately represent all content areas and cognitive processes specified in the test blueprint?

Efficiency: This refers to the number of items per unit of testing time. The more information about a student's achievement level obtained in a specific amount of time, the better.

Specificity: Items should be written to measure learning objectives only, not reading or writing ability, general intelligence, or test taking ability.

Difficulty

*Norm-referenced:* A difficulty level represents the proportion of examinees responding correctly to an item. Measurement specialists suggest an ideal mean difficulty for a norm-referenced achievement test to be halfway between a perfect score and a chance score. For example, if there are four response options, a chance score is 25% and 62.50 is the ideal average difficulty. Also, measurement experts believe that four-option multiple-choice items with difficulty levels below .5 (less than 50% passing) are too difficult. Either there is a problem with the item itself or the content is not understood. Another possibility is that students are accustomed to studying for multiple-choice tests written at the rote level, and may not be prepared for a test requiring higher cognitive levels. Thus, provide students with examples of the types of items the test will include.

*Criterion-referenced*: For minimum-competency criterion-referenced tests, because a large proportion of examinees should answer correctly, the same average criterion does not apply. The difficulty for a criterion-referenced test should be consistent with the logically or empirically based predetermined criterion. For example, if 80% is the criterion identified, the difficulty levels should be similar to that percentage.

Discrimination

*Norm-referenced*: This index shows how well items discriminate between the high and low achieving students. Discrimination indices range from –1.00 to +1.00, with a positive index indicating that students who performed well on the test tended to answer the item correctly. A negative discriminator, suggesting that the poorly performing students tended to answer the item correctly, is undesirable.

*Criterion-referenced*: Since the discrimination index included in the item analysis printout provided by Evaluation and Testing is based upon correlation, which is dependent upon variability, the discrimination index would be expected to be low for mastery-type items.

Variability

*Norm-referenced*: If grades are to be based upon a normal curve, a wide range or spread of scores is necessary. Very easy or difficult items do not contribute to variability.

*Criterion-referenced*: If mastery is being evaluated, wide variation in test scores is not expected or desired. The resulting distribution should be negatively skewed rather than normally distributed.

Reliability

*Norm-referenced*: This is an overall test statistic indicating score consistency, and is the single most important statistic for a norm-referenced achievement test. Reliability indices range from 0.0 to 1.00 with a .70 considered to be the minimum value acceptable. High score variability, high discrimination, and moderate difficulty levels are associated with high reliability. Although necessary if grades are to be assigned and to have confidence in the scores, reliability does not ensure validity or relevance.

*Criterion-referenced*: Most of the contributors to high reliability are not sought for criterion-referenced tests. It was previously suggested that the standard error of measurement could be used to evaluate the consistency of scores for either type of test and is the preferred statistic to measure stability of results from criterion-referenced tests. Bear in mind that validity, which is represented by Relevance and Balance in this discussion, is the most important criterion for either norm or criterion-referenced tests.

# USE OF THE ITEM ANALYSIS TO IMPROVE ITEMS, TESTS AND TEACHING

This section explains the information provided in the item analysis printout produced by Evaluation and Testing. Item evaluations and interpretations are presented from both the norm-referenced and criterion-referenced perspectives. Refer to Appendix D when reading.

**Reading and Using the Item Analysis Printout**

The item analysis printout begins with a frequency distribution, and includes the following information: the scores earned ("Score"); the frequency of each score ("Freq"); the cumulative frequency ("Cum F") starting with the lowest score and summing the frequency from the lowest to the specific score; the proportion of students earning a particular score ("PRP"); the cumulative proportion ("Cum P"); and "Z" score. Each score has been transformed into a Z, or standard score, which ranges from –3.00 to +3.00. The Z score can be compared to a standard normal curve to determine the percentile. The mean, median, or point below which 50% of the scores fall, standard deviation ("Std. Dev."), which is roughly the average variability around the mean, and the number of cases ("N") are listed at the bottom of the printout.

The actual item analysis follows the frequency distribution. The "Item No.", located in the first and last columns (Appendix D, Pg. 2) identifies the item. The second column, "Key" presents the correct response as indicated by the answer key. The headings "1-5" refer to the specific response options, both the correct response and the distractors. Numbers will be in the "other" column only if students gridded more than one option for that specific item. The "N" and "P" under each number represents the number and proportion of students selecting that particular response option. The "Pro.

Passing" column contains the difficulty value for the item, with numbers close to 1.00 indicating that the majority of students answered correctly. The "D" column lists the discrimination index, which can range from –1.00 to +1.00. Positive indices suggest that the better performing students tended to answer the item correctly.

There are also important data at the bottom of the last page: 1) "N", the number of examinees tested, 2) "Mean", the arithmetic average for the test, 3) "St. Dev.", the standard deviation, 4) "KR-20", the reliability estimate, which can range from 0.0 to +1.00, 5) "S.E.", the standard error of measurement, which estimate the error attributed to the test by comparing the reliability estimates to the standard deviation.

A student's theoretical "true score" is the summation of the observed or actual score and error score. By using these two values, one can estimate the "true score" with a certain degree of accuracy. The standard error of measurement can be related to the Z score and a student's score to determine the range of possible "true scores" with the formula: "True Score = X $\pm$ Z x S.E." For example, if the S.E. is 3 and the student's score is 50, the "true score" would fall between 50 $\pm$ (1) 3, or 47-53, 68% of the time since $\pm$ (1) Z encompasses 68% of the standard normal distribution. The "true score" should be within the range of 50 $\pm$ (2) 3, or 44-56, 95% of the time based on the probabilities associated with the standard normal curve, because 95% of the distribution falls within two standard deviations. The "true score" range within three standard deviations of the mean is 50 $\pm$ (3) 3, or 41-59, which would capture the "true score" 99 times out of 100. This statistic and method can be used to judge the precision of scores for either criterion-referenced or norm-referenced tests.

Alternatively, the size of the standard error of measurement can be evaluated by itself to judge the precision of scores for either criterion-referenced or norm-referenced tests.

Please note that the impact of an atypical score on any statistic is greater when the statistic is calculated from a small set of numbers ($n < 25$) than when the statistic is calculated from a large set of numbers.

**Using the Item Analysis Printout to Evaluate the Test**

For an achievement test to be useful for assigning grades, the table of item specifications representing the course content and appropriate cognitive levels must be followed. The test must also produce consistent results, and therefore have adequate reliability estimates. Items and instruction can be evaluated by using the difficulty index, discrimination index, and the distractor analysis.

Begin the evaluation of the test and items by reviewing the distribution of scores. If a test has been written to identify individual differences, a normal or bell-shaped curve or a flatter version of this curve is expected. If a criterion-referenced test or one which includes mastery items has been written, a skewed distribution with the majority of scores in the upper end with fewer scores on the negative end of the distribution is anticipated. Because variability is typically lower for the criterion-referenced exam, lower reliability and discrimination indices may not suggest problems with the items on a criterion-referenced test as they would with a norm-referenced test which should have wider score variation.

The reliability coefficient (KR-20) is evaluated next, with a .70 or higher value expected if a norm-referenced test has been developed. Since mastery-type tests should have less variation, the reliability coefficient will be lower.

An inspection of the data presented in Appendix D reveals a KR-20 of .754, an acceptable statistic, but expected given the number of items on the test (n = 100). (Refer to Addressing Reliability.)

After reviewing the distribution and reliability estimate (KR-20) of the items, the difficulty index, the discrimination index and the distractors should be scrutinized.

*Difficulty Index (Proportion Passing) Analysis*

The difficulty index, the proportion of students answering an item correctly, can range from 0.0 to +1.00. A D = 0.0 indicates that not one student answered the item correctly and the item was very difficult, ambiguous, or miskeyed. When 100% of the examinees answer an item correctly (D = 1.00), either clues indicating the correct response were given or the content was very basic and was mastered. Measurement experts maintain that an item with a difficulty index under .50 is too difficult. A low index may occur because the content was complex, the item was faulty, instruction was inadequate or students were not prepared. In addition to the difficulty level for each item, the average difficulty for a norm-referenced test should be located halfway between a chance score and a perfect score (100%). Thus, for a four-option test, the average difficulty should be 62.5 and is calculated in the following manner:

$$\frac{(100 - 25)}{2} = 37.5$$ (100% = perfect score, 25% = chance score – 25 of 100)

$37.5 + 25 = 62.5$ (added to the chance score)

The average difficulty calculated for the printout in Appendix D is approximately .71, which is higher than the recommended difficulty for a norm-referenced test. An observed difficulty that is much higher or lower than the recommended difficulty could have a varying negative impact on the reliability index because moderate difficulty adds to variability, which is a contributor to the reliability of a norm-referenced test. A very high average difficulty (very easy items) for a norm-referenced test could indicate a poorly written test, an easy test, or one which provides clues to the correct responses. However, if a criterion-referenced test has been administered, high average difficulty is expected. A higher than average difficulty may represent a test which is not strictly norm-referenced, which may or may not have been the instructor's intent. The test may be truly reflective of course content and the table of item specifications but does not have the identification of individual differences or gaining the full range of achievement as its primary purpose. Another possibility is that the instructor did not adhere to acceptable test construction procedures.

*Item Discrimination Analysis*

The item discrimination index compares students' performance on an item to their performance on the entire examination. The point-biserial correlation index, the method used by Evaluation and Testing to measure item discrimination, compares the performance of all students on each item to their performance on the total test. Another method of deriving an item discrimination index compares examinees' performance in the upper and lower groups on the examination (e.g. upper and lower 27%) for each item. Item discrimination indices vary from –1.00 to +1.00, with a negative index suggesting that poorly performing students on the exam answered the particular item

correctly, or conversely, high performing students answered the item incorrectly. Items should have a positive discrimination index, indicating that those who scored high on the test also tended to answer the item correctly, while those who scored low on the test tended to answer incorrectly.

A discrimination index should be evaluated with reference to the difficulty level of the item (see Item Analysis Summary), because a correlation method is used to assess the item's success in discriminating between low and high achieving students. If the items are very easy or difficult, indicating homogenous performance, there is less variation in the scores, thus resulting in a reduced potential for discrimination. For example, item #31 has a slightly negative discrimination index, but is an extremely difficult item. Therefore, at least one poorly performing student answered this very difficult item correctly. This item could be evaluated as being too difficult and having ambiguous options and should be revised. It is also possible that the correct option was miskeyed or the content was not taught as thoroughly as the instructor thought. Regardless of the reason for this result, negative discrimination indices are undesirable.

Item #8 is a negative discriminator, with nearly 20% of the students selecting the incorrect response. This could be a result of an ambiguously worded stem or options, or some students who performed poorly on the test answering this item correctly. Several items on this exam are relatively to very difficult (prop. passing < .50). In most cases, although difficult, the discrimination index (D) is relatively strong (.2 and above).

The discrimination index can be used to evaluate an item from either a criterion-referenced test or a norm-referenced achievement test, however, less variation is expected on the criterion-referenced test. Students who were successful in mastering

the material overall should answer the items correctly, resulting in a positive but lower discrimination index.

*Distractor Analysis*

It was mentioned in the Developing Items section that distractors must be plausible. This suggests that at least for norm-referenced tests, some students should be attracted to every distractor. Ideally, at least one person should select each one of the incorrect options. If this does not happen, the options should be reviewed for plausibility. In a criterion-referenced test, it is also necessary to be assured that the criterion has been attained. Therefore, distractors from these tests also should be plausible.

*Item Analysis Summary*

The three components of an item analysis are interdependent. One approach to item evaluation begins with reviewing the difficulty level. An instructor should have expectations about how students will perform on the item regardless if a norm-referenced or criterion-referenced test has been written. If the item is expected to be easy for most students, and it is not, the discrimination index should be reviewed. If the index is positive, indicating that the better performing students answered the item correctly, then the content covered on the item may not have been not fully understood by the students. This suggests that further instruction is needed, not that the item was poor. If the discrimination index is negative, the item needs revision. When the content is complex, a low difficulty index (few answering the item correctly) may be expected unless the class is composed of a homogeneous group of high achievers who have thoroughly grasped the content. If most students correctly answered an item expected

to be very difficult, the correct response option could contain clues or the distractors could be implausible. At this point, the distractors should be reviewed.

## TABLE 6

| Item # | Key | 1 | | 2 | | 3 | | 4 | | 5 | | Other | | Prop. Passing | Mean of Passers | D | Item # |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | N | P | N | P | N | P | N | P | N | P | N | P | | | | |
| 1 | 1 | 21 | 1.00 | 0 | .00 | 0 | .00 | 0 | .00 | 0 | .00 | 0 | .00 | 1.00 | 74.5 | .000 | 1 |
| 3 | 4 | 2 | .10 | 2 | .10 | 0 | .00 | 17 | .81 | 0 | .00 | 0 | .00 | .81 | 77.8 | .609 | 3 |
| 4 | 2 | 0 | .00 | 18 | .86 | 3 | .14 | 0 | .00 | 0 | .00 | 0 | .00 | .86 | 76.8 | .510 | 4 |
| 8 | 3 | 0 | .00 | 1 | .05 | 17 | .81 | 2 | .10 | 1 | .05 | 0 | .00 | .81 | 73.5 | -.183 | 8 |
| 9 | 2 | 2 | .10 | 11 | .52 | 5 | .24 | 3 | .14 | 0 | .00 | 0 | .00 | .52 | 77.7 | .301 | 9 |
| 11 | 2 | 3 | .14 | 18 | .86 | 0 | .00 | 0 | .00 | 0 | .00 | 0 | .00 | .86 | 75.9 | .317 | 11 |
| 17 | 3 | 4 | .19 | 5 | .24 | 10 | .48 | 0 | .00 | 2 | .10 | 0 | .00 | .48 | 75.8 | .111 | 17 |
| 35 | 2 | 4 | .19 | 15 | .71 | 1 | .05 | 0 | .00 | 1 | .05 | 0 | .00 | .71 | 78.0 | .492 | 35 |
| 40 | 3 | 0 | .00 | 1 | .05 | 20 | .95 | 0 | .00 | 0 | .00 | 0 | .00 | .95 | 74.7 | .088 | 40 |
| 97 | 1 | 6 | .29 | 1 | .05 | 0 | .00 | 10 | .48 | 4 | .19 | 0 | .00 | .29 | 84.0 | .532 | 97 |

Another item evaluation approach begins with an analysis of the discrimination index, followed by a review of the difficulty with an and finally an inspection of the item options, if warranted. Items from Appendix D (see Table 6 above) were selected to explain this process from both the norm-referenced and criterion-referenced perspectives. Item #1 has no discrimination ability (D = 0), with 100% of the examinees answering correctly (Pro. Passing = 1.00). This indicates an easy item, either due to content mastery or clues in the correct response. The instructor was probably trying to build examinees' confidence. Since Item #4 was answered correctly by the majority of examinees and was a strong positive discriminator, only the most poorly performing students answered incorrectly. The discrimination index and proportion passing suggest the item is performing properly and is appropriate for either testing purpose, assuming the test blueprint was followed carefully. It is acceptable to begin even a

norm-referenced test with easier items. Item #40 has a low positive discrimination index and a high proportion passing. Although the discrimination potential is limited, the better students appeared to answer the item correctly. This item is satisfactory if the instructor was evaluating a rudimentary fact or concept which should be thoroughly understood. Item #35 is a strong positive discriminator with moderate difficulty. This is an excellent item for measuring individual differences. However, if these results are obtained on a mastery-type item, a higher passing rate may be expected. If a higher rate was anticipated, additional instruction may be needed. Item #97 has a positive discrimination index, but a low proportion passing. Because the item was a positive discriminator, it is probable that this item is satisfactory, although difficult. These results suggest the need for additional instruction since many students do not understand this material. Items #9 and #17 could be evaluated similarly. While being positive discriminators, approximately 50% of the students answered incorrectly, indicating they did not fully grasp the content. Alternatively, they did not expect this type of multiple choice test, and therefore did not study adequately. An inspection of the item and test blueprint should provide information regarding the plausibility of these explanations. Ideally, the proportion passing should be higher. From either a norm-referenced or criterion-referenced perspective, additional instruction may be indicated. Item #11 is a positive discriminator, with a larger proportion of students answering the item correctly than incorrectly. This is probably a good item for either testing purpose. The results from Item #3 reveal a majority of students passing and a positive discrimination index. This is probably a good item for either a criterion-referenced or norm-referenced exam. If the table of item specifications calls for an item with a large percentage of students

answering correctly, then this item is suitable for a norm-referenced test. Item #8 is a negative discriminator, indicating the better performing students did not answer the item correctly. This is undesirable for either a criterion-referenced or norm-referenced test. If an item is moderately difficult, a strong and positive discrimination index is expected from a norm-referenced test. If the discrimination index is low and the item is moderately difficult, typically the item should be revised. Determine if the correct response is ambiguous, if the option or answer was confusing, or if the material was not understood.

Very easy or very difficult items are not expected to have high discrimination indices, since the discrimination index is based upon correlation. Because a strong correlation requires variation in the distribution, with little variation the potential for discrimination is reduced. Thus, when the discrimination index is low and the item is very easy or difficult, there is less concern than when an item with a low discrimination index (index close to 0.0) or moderate difficulty are observed.

To summarize, each item should be evaluated within the framework of the type of test, norm-referenced or criterion-referenced. A general level of difficulty should be expected for each item and should be compared to the observed difficulty. Positive discrimination indices (.2 and above) are desired. If a problem is uncovered, evaluate the options and decide if the item and/or the instruction need revision.

*Synthesis*

There has been considerable discussion about the purposes and interpretations of criterion-referenced and norm-referenced achievement examinations. There are commonalties in the two evaluation approaches. In fact, an inspection of the

development stages may not reveal which evaluation approach has been used, because both should begin with a table of item specifications, which reflects the course content and learning outcomes. The two types of tests have different purposes; the criterion-referenced approach evaluates how well each student performs in relation to a predetermined criterion and the norm-referenced approach compares students in relation. The appropriate approach to testing should be based upon instructional goals. If an exact criterion can be established, for example, if 80% of the items on the test represent mastery of specific content, and the achievement of this criterion is critical, then criterion-referenced testing is the appropriate vehicle to use. If students will be ranked based upon their achievement of content and learning objectives, then a norm-referenced test which emphasizes items that maximize these differences is appropriate.

To develop a norm-referenced test which will reliably identify individual differences, construct and adhere to a table of item specifications and write items that are moderately difficult and positive discriminators. However, if content is covered that is fundamental, the difficulty index should be high (majority of students responded correctly), and the discrimination index should be positive, but may be low. If testing a homogeneous group of high achievers, such as an advanced graduate class, the range of performance on any item or the entire test should be restricted, with a large proportion of students passing each item. For this class, relatively low discrimination indices may result. While it may be necessary to measure individual differences in student achievement for this group of students, the items and scores may reflect a distribution more typical of a criterion-referenced exam than of a norm-referenced exam.

To summarize, the main differences between criterion-referenced and norm-referenced tests occur at the criterion-setting point and at the item analysis, test statistic review point. The table of item specifications guides the development of both the norm-referenced and criterion-referenced tests.

# APPENDIX A
## Condensed Version of the Taxonomy of Educational Objectives

## Cognitive Domain

### *KNOWLEDGE*

### 1.00 KNOWLEDGE

Knowledge, as defined here, involves the recall of specifics and universals, the recall of methods and processes, or the recall of a pattern, structure, or setting. For measurement purposes, the recall situation involves little more than bringing to mind the appropriate material. Although some alteration of the material may be required, this is a relatively minor part of the task. The knowledge objectives emphasize most the psychological processes of remembering. The process of relating is also involved in that a knowledge test situation requires the organization and reorganization of a problem such that it will furnish the appropriate signals and cues for the information and knowledge the individual possesses. To use an analogy, if one thinks of the mind as a file, the problem in a knowledge test situation is that of finding in the problem or task the appropriate signals, cues, and clues which will most effectively bring out whatever knowledge is filed or stored.

### 1.10 KNOWLEDGE OF SPECIFICS

The recall of specific and isolated bits of information. The emphasis is on symbols with concrete referents. This material, which is at a very low level of abstraction, may be thought of as the elements from which more complex and abstract forms of knowledge are built.

## 1.11 KNOWLEDGE OF TERMINOLOGY

Knowledge of the referents for specific symbols (verbal and non-verbal). This may include knowledge of the most generally accepted symbol referent, knowledge of the variety of symbols which may be used for a single referent, or knowledge of the referent most appropriate to a given use of a symbol.

- To define technical terms by giving their attributes, properties, or relations.
- Familiarity with a large number of words in their common range of meanings.

## 1.12 KNOWLEDGE OF SPECIFIC FACTS

Knowledge of dates – events, persons, places, etc. This may include very precise and specific information such as the specific date or exact magnitude of a phenomenon. It may also include approximate or relative information such as an approximate time period or the general order of magnitude of a phenomenon.

- The recall of major facts about particular cultures.
- The possession of a minimum knowledge about the organisms studied in the laboratory.
- Illustrative education objectives selected from the literature.

## 1.20 KNOWLEDGE OF WAYS AND MEANS OF DEALING WITH SPECIFICS

Knowledge of the ways of organizing, studying, judging, and criticizing. This includes the methods of inquiry, the chronological sequences, and the standards of judgement within a field as well as the patterns of organization through which the areas of the fields themselves are determined and internally organized. This knowledge is at an intermediate level of abstraction between specific knowledge on the one hand and knowledge of universals on the other. It does not so much demand the activity of the student in using the materials as it does a more passive awareness of their nature.

## 1.21 KNOWLEDGE OF CONVENTIONS

Knowledge of characteristic ways of treating and presenting ideas and phenomena. For purposes of communication and consistency, workers in a field employ usages, styles, practices and forms which best suit their purposes and/or which appear to suit best the phenomena with which they deal. It should be recognized that although these forms and conventions are likely to be set up on arbitrary, accidental, or authoritative bases, they are retained because of the general agreement or concurrence of individuals concerned with the subject, phenomena, or problem.

- Familiarity with the forms and conventions of the major types of works, e.g., verse, plays, scientific papers, etc.
- To make pupils conscious of correct form and usage in speech and writing.

## 1.22 KNOWLEDGE OF TRENDS AND SEQUENCES

Knowledge of the processes, directions, and movements of phenomena with respect to time.

- Understanding of the continuity and development of American culture as exemplified in American life.
- Knowledge of the basic trends underlying the development of public assistance programs.

## 1.23 KNOWLEDGE OF CLASSIFICATIONS AND CATEGORIES

Knowledge of the classes, sets, divisions, and arrangements which are regarded as fundamental for a given subject field, purpose, argument, or problem.

- To recognize the area encompassed by various kinds of problems or materials.
- Becoming familiar with a range of types of literature.

## 1.24 KNOWLEDGE OF CRITERIA

Knowledge of the criteria by which facts, principles, opinions, and conduct are tested or judged.

- Familiarity with criteria for judgement appropriate to the type of work and the purpose for which it is read.
- Knowledge of criteria for the evaluation of recreational activities.

## 1.25 KNOWLEDGE OF THE METHODOLOGY

Knowledge of the methods of inquiry, techniques, and procedures employed in a particular subject field as well as those employed in investigating particular problems and phenomena. The emphasis here is on the individual's knowledge of the method rather than his ability to use the method.

- Knowledge of scientific methods for evaluating health concepts.
- The student shall know the methods of attack relevant to the kinds of problems of concern to the social sciences.

## 1.30 KNOWLEDGE OF THE UNIVERSALS AND ABSTRACTIONS IN A FIELD

Knowledge of the major schemes and patterns by which phenomena and ideas are organized. These are the large structures, theories, and generalizations which dominate a subject field or which are quite generally used in studying phenomena or solving problems. These are at the highest levels of abstraction and complexity.

## 1.31 KNOWLEDGE OF PRINCIPLES AND GENERALIZATIONS

Knowledge of particular abstractions which summarize observations of phenomena. These are the abstractions which are of value in explaining, describing, predicting, or in determining the most appropriate and relevant action to be taken.

- Knowledge of the important principles by which our experience with biological phenomena is summarized.
- The recall of major generalizations about particular cultures.

### 1.32 KNOWLEDGE OF THEORIES AND STRUCTURES

Knowledge of the body of principles and generalizations together with their interrelations which present a clear, rounded, and systematic view of a complex phenomenon, problem, or field. These are the most abstract formulations, and they can be used to show the interrelation and organization of a great range of specifics.

- The recall of major theories about particular cultures.
- Knowledge of a relatively complete formulation of the theory of evolution.

## ***INTELLECTUAL ABILITIES AND SKILLS***

Abilities and skills refer to organized modes of operation and generalized techniques for dealing with materials and problems. The materials and problems may be of such a nature that little or no specialized and technical information is required. Such information as is required can be assumed to be part of the individual's general fund of knowledge. Other problems may require specialized and technical information at a rather high level such that specific knowledge and skill in dealing with the problem and the materials are required. The abilities and skills objectives emphasize the mental processes of organizing and reorganizing material to achieve a particular purpose. The materials may be given or remembered.

### 2.00 COMPREHENSION

This represents the lowest level of understanding. It refers to a type of understanding or comprehension such that the individual knows what is being communicated and can make use of the material or idea being communicated without necessarily relating it to other material or seeing its fullest implications.

## 2.10 TRANSLATION

Comprehension as evidenced by the care and accuracy with which the communication is paraphrased or rendered from one language or form of communication to another. Translation is judged on the basis of faithfulness and accuracy, that is, on the extent to which the material in the original communication is preserved although the form of the communication has been altered.

- The ability to understand non-literal statements (metaphor, symbolism, irony, exaggeration).
- Skill in translating mathematical verbal material into symbolic statements and vice versa.

## 2.20 INTERPRETATION

The explanation or summarization of a communication. Whereas translation involves an objective part-for-part rendering of a communication, interpretation involves a reordering, rearrangement, or a new view of the material.

- The ability to grasp the thought of the work as a whole at any desired level of generality.
- The ability to interpret various types of social data.

## 2.30 EXTRAPOLATION

The extension of trends or tendencies beyond the given data to determine implications, consequences, corollaries, effects, etc.; which are in accordance with the conditions described in the original communication.

- The ability to deal with the conclusions of a work in terms of the immediate inference made from the explicit statements.
- Skill in predicting continuation of trends.

## 3.00 APPLICATION

The use of abstractions in particular and concrete situations. The abstractions may be in the form of general ideas, rule or procedures, or generalized methods. The

abstractions may also be technical principles, ideas, and theories which must be remembered and applied.

- Application to the phenomena discussed in one paper of the scientific terms or concepts used in other papers.
- The ability to predict the probable effect of a change in a factor on a biological situation previously at equilibrium.

**4.00 ANALYSIS**

The breakdown of a communication into its constituent elements or parts such that the relative hierarchy of ideas is made clear and/or the relations between the ideas expressed are made explicit. Such analyses are intended to clarify the communication, to indicate how the communication is organized, and the way in which it manages to convey its effects, as well as its basis and arrangement.

**4.10 ANALYSIS OF ELEMENTS**

Identification of the elements included in a communication.

- The ability to recognize unstated assumptions.
- Skills in distinguishing facts from hypotheses.

**4.20 ANALYSIS OF RELATIONSHIPS**

The connections and interactions between elements and parts of a communication.

- Ability to check the consistency of hypotheses with given information and assumptions.
- Skill in comprehending the interrelationships among the ideas in a passage.

**4.30 ANALYSIS OF ORGANIZATIONAL PRINCIPLES**

The organization, systematic arrangement, and structure which hold the communication together. This includes the “explicit” as well as “implicit” structure. It

includes the bases, necessary arrangement, and the mechanics which make the communication a unit.

- The ability to recognize form and pattern in literary or artistic works as a means of understanding their meaning.
- Ability to recognize the general techniques used in persuasive materials, such as advertising, propaganda, etc.

**5.00 SYNTHESIS**

The putting together of elements and parts so as to form a whole. This involves the process of working with pieces, parts, elements, etc., and arranging and combining them in such a way as to constitute a patter or structure not clearly there before.

**5.10 PRODUCTION OF A UNIQUE COMMUNICATION**

The development of a communication in which the writer or speaker attempts to convey ideas, feelings, and/or experiences to others.

- Skill in writing, using an excellent organization of ideas and statements.
- Ability to tell a personal experience effectively.

**5.20 PRODUCTION OF A PLAN, OR PROPOSED SET OF OPERATIONS**

The development of a plan of work or the proposal of a plan of operations. The plan should satisfy requirements of the task which may be given to the student or which he may develop for himself.

- Ability to propose ways of testing hypotheses.
- Ability to plan a unit of instruction for a particular teaching situation.

**5.30 DERIVATION OF A SET OF ABSTRACT RELATIONS**

The development of a set of abstract relations either to classify or explain particular data or phenomena, or the deduction of propositions and relations from a set of basic propositions or symbolic representations.

- Ability to formulate appropriate hypotheses based upon an analysis of factors involved, and to modify such hypotheses in the light of new factors and considerations.
- Ability to make mathematical discoveries and generalizations.

## 6.00 EVALUATION

Judgements about the value of material and methods for given purposes. Quantitative and qualitative judgements about the extent to which material and methods satisfy criteria. Use of a standard of appraisal. The criteria may be arranging and combining them in such a way as to constitute a patter or structure not clearly there before.

## 6.10 JUDGEMENTS IN TERMS OF INTERNAL EVIDENCE

Evaluation of the accuracy of a communication from such evidence as logical accuracy, consistency, and other internal criteria.

- Judging by internal standards, the ability to assess general probability of accuracy in reporting facts from the care given to exactness of statement, documentation, proof, etc.
- The ability to indicate logical fallacies in arguments.

## 6.20 JUDGEMENTS IN TERMS OF EXTERNAL CRITERIA

Evaluation of material with reference to selected or remembered criteria.

- The comparison of major theories, generalizations, and facts about particular cultures.
- Judging by external standards, the ability to compare a work with the highest known standards in its field – especially with other works of recognized excellence.

# APPENDIX B
## Correct Responses and Clues for Activity 1

| Question | Correct Response | Clue |
|---|---|---|
| 1 | A | This option is much longer than the other options. Also, a pact is a formal agreement usually between nations. |
| 2 | C | This option is the only option grammatically correct with the stem. |
| 3 | A | This option repeats information in the stem. |
| 4 | A | This is the only option without the word “every”. Typically “every” and other absolute words are incorrect. |
| 5 | D | Question 1 provides information about the stem. |
| 6 | A | Options B and C are similar and A is dissimilar, making Option D incorrect. Also, B and C are incorrect because they are similar. |
| 7 | C | This is the only option that is specific. |

# APPENDIX C
## Bloom's Taxonomy of Educational Objectives

I. KNOWLEDGE

A. Remembers by recall or recognition; should not be too different from way in which knowledge was originally learned.

B. Behavior tasks

1. Defines…
2. Recalls…
3. Lists…
4. States…
5. Recites…
6. Names…
7. Describes…
8. Selects…

II. COMPREHENSION

A. Grasps the meaning of the material; deals with the content. Requires interpretation or translation from abstract to simple phrases or making generalizations.

B. Behavioral tasks

1. States in own words…
2. Gives an example of…
3. Illustrates…
4. Summarizes…
5. Interprets…
6. Classifies…
7. Explains…
8. Predicts…
9. Distinguishes between…

III. APPLICATION

A. Uses information in real-life problems.

B. Behavioral tasks

1. Chooses appropriate procedure…
2. Applies a principle…
3. Uses an approach…
4. Solves a problem…
5. Computes…
6. Relates…
7. Demonstrates…

IV. ANALYSIS

A. Breaks material into constituent parts and identifies relationships of the parts to each other and to the whole.

B. Distinguishes fact from hypothesis and from value statements.

C. Identifies conclusions and generalizations.

D. Separates relevant from trivia.

E. Differentiates one symbol from another symbol.

F. Behavioral tasks

1. Distinguishes…
2. Discriminates between…
5. Differentiates
6. Infers…

3. Discovers…
4. Detects…
7. Subdivides…

V. SYNTHESIS

A. Combines parts to make a whole.
B. Emphasizes originality.
C. Organizes ideas into new patterns.
D. Tries various approaches.
E. Ability to use results of research in solving a problem.
F. Behavioral tasks
1. Develops…
2. Writes…
3. Designs…
4. Creates…
5. Combines…
6. Composes…

VI. EVALUATION

A. Makes a judgement concerning the value of ideas, principles, methods, solutions, etc.
B. Uses set criteria.
C. Not opinions.
D. Recognizes fallacies.
E. Behavior tasks
1. Compares…
2. Judges…
3. Determines the best possible…
4. Applies criteria…
5. Concludes…
6. Discriminates…
7. Supports…

*Adapted from Constructing Achievement Tests, Norman E. Gronlund and Workshops presented by Bruce Hall, University of South Florida.*

# BIBLIOGRAPHY


Bloom, Benjamin S. 1956 Ed., Taxonomy of Educational Objectives, David McKay Company, Inc., New York.

Carey, Lou M. 1988, Measuring and Evaluating School Learning, Allyn and Bacon, Inc., Newton, MA.

Crocker, Linda and Algina, James. 1986, Introduction to Classical and Modern Test Theory, CBS College Publishing, New York.

Ebel, Robert L. and Frisbie, David A. 1991, Essentials of Educational Measurement, 5th ed., Prentice-Hall, Englewood Cliffs, NJ.

Ebel, Robert L. 1980, Practical Problems in Educational Measurement, D.C. Health and Co., Lexington, MA.

Eison, James. 1985, Constructing Classroom Tests: Why and How, Southeast Missouri State University, Cape Girardeau, MO.

Gronlund, Norman E. 1982, Constructing Achievement Tests, 3rd ed., Prentice-Hall, Inc., Englewood Cliffs, NJ.

Gronlund, Norman E. 1993, How to Make Achievement Tests and Assessments, 5th ed., Allyn and Bacon, Needham Heights, MA.

Hall, Bruce. 1993, Handouts distributed in workshops, University of South Florida, Tampa, FL. Permission granted to use materials.

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

Hopkins, K.D., Stanley, J.C. and Hopkins, B.R. 1990, Educational and Psychological Measurement and Evaluation, 7th ed., Prentice-Hall, Englewood Cliffs, NJ.

Sax, Gilbert. 1989, Principles of Educational and Psychological Measurement and Evaluation, 3rd ed., Wadsworth, Inc., Belmont, CA.

Zimmerman, Beverly B., Sudweeks, R.R., Shelley, M.F., Wood, Bud. 1990, How to Prepare Better Tests: Guidelines for University Faculty, Brigham Young University Testing Services and the Department of Instructional Science, Provo, UT.