Writing Good Multiple-Choice Exams

Dawn M. Zimmaro, Ph.D.

Measurement and Evaluation Center
Telephone: (512) 232-2662
Web: www.utexas.edu/academic/mec
Location: Bridgeway Building, 2616 Wichita Street
Address: P.O. Box 7246, Austin, TX 78713-7246

DIIA
THE UNIVERSITY OF TEXAS AT AUSTIN the division of instructional innovation and assessment • www.utexas.edu/academic/diia

Last revised August 19, 2004

Table of Contents

Goals of the Workshop

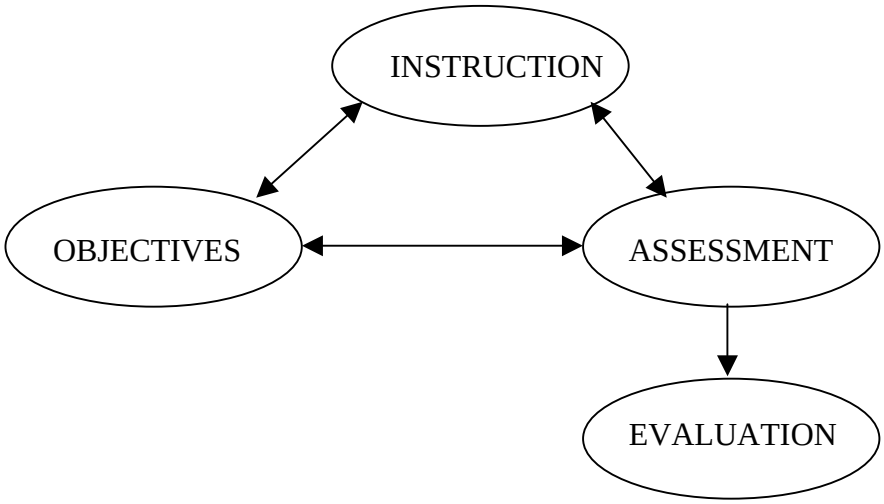
Multiple-choice exams are commonly used to assess student learning. However, instructors often find it challenging to write good items that ask students to do more than memorize facts and details. In this workshop we will explore how to create effective classroom multiple-choice exams that are based on sound learning objectives and how you can use information from your exams to improve your teaching.

After completing the workshop you will be able to:

- Describe various levels of learning objectives
- Explain the strengths and weaknesses of multiple-choice exams
- Identify common errors when writing multiple-choice items
- Create multiple-choice items that assess various levels of learning
- Use exam results for feedback and to evaluate instructional effectiveness

The KEY to Effective Testing

- To maximize your testing, you should aim to integrate all the major components of a course.



OBJECTIVES: Specific statements of the goals of the instruction; the objectives express what the students should be able to do or know as a result of taking the course; also, the objectives should indicate the cognitive level of performance expected (e.g., basic knowledge level, deeper comprehension level, or application level).

INSTRUCTION: This consists of all the usual elements of the curriculum designed to teach a course, including lesson plans, study guides, and reading and homework assignments; the instruction should correspond directly to the course objectives.

ASSESSMENT: The process of gathering, describing, or quantifying information about performance; the testing component of the course; the amount of weight given to the different subject matter areas on the test should match the relative importance of each of the course objectives as well as the emphasis given too each subject area during instruction.

EVALUATION: Examining student performance and comparing and judging its quality. Determining whether or not the learner has met the course objectives and how well.

Summary of How Evaluation, Assessment, Measurement and Testing Terms Are Related

Commonly used assessment and measurement terms are related and understanding how they connect with one another can help you better integrate your testing and teaching.

Evaluation

Examining information about many components of the thing being evaluated (e.g., student work, schools, or a specific educational program) and comparing or judging its quality, worth or effectiveness in order to make decisions

based on ↓

Assessment

The process of gathering, describing, or quantifying information about performance.

includes ↓

Measurement

Process of assigning numbers to qualities or characteristics of an object or person according to some rule or scale and analyzing that data based on psychometric and statistical theory

specific way to measure performance is ↓

Testing

A method used to measure the level of achievement or performance

Course Learning Objectives

Course objectives should contain clear statements about what the instructor wants to know by the end of the semester. If objectives are clearly and specifically defined, the instructor will have an effective means of evaluating what the students learned.

Course objectives should not be so specific that the creativity of the instructor and student are stifled, nor should they be so vague that the students are left without direction.

An example of a well constructed objective might be: "Students in Psychology 100 will be able to demonstrate their knowledge of Erikson's Psychosocial Stages of Development by naming the 8 stages in order and describing the psychosocial crises at each stage."

Note that the objective is written in terms of what the ***student*** will be able to do, not what the instructor will teach. Learning objectives should focus on what the students should be able to do or know at the end of the semester.

Do not use words that can be open to interpretation or are unclear. "Students should have an understanding of Erikson's theory of development." How would you measure "an understanding" or "an awareness" or "an appreciation"?

In beginning to write course learning objectives you may find it helpful to write some general statements about some concepts, topics, and principles of course content. From those general statements you can then write specific objectives for class sessions.

Bloom specified different abilities and behaviors that are related to thinking processes in his Taxonomy of Educational Objectives. This taxonomy can be helpful in outlining your course learning objectives.

Reference:

Hellyer, S. (n.d.). *A teaching handbook for university faculty. Chapter 1: Course objectives.* Retrieved October 1, 1998 from Indiana University Purdue University Indianapolis Web site: http://www.iupui.edu/~profdev/handbook/chap1.html

Abilities and Behaviors Related to Bloom's Taxonomy of Educational Objectives

Knowledge – Recognizes students' ability to use rote memorization and recall certain facts.

- Test questions focus on identification and recall of information

Comprehension – Involves students' ability to read course content, extrapolate and interpret important information and put other's ideas into their own words.

- Test questions focus on use of facts, rules and principles

Application – Students take new concepts and apply them to another situation.

- Test questions focus on applying facts or principles

Analysis – Students have the ability to take new information and break it down into parts to differentiate between them.

- Test questions focus on separation of a whole into component parts

Synthesis – Students are able to take various pieces of information and form a whole creating a pattern where one did not previously exist.

- Test questions focus on combining ideas to form a new whole

Evaluation – Involves students' ability to look at someone else's ideas or principles and see the worth of the work and the value of the conclusions.

- Test questions focus on developing opinions, judgments or decisions

Reference:

Hellyer, S. (n.d.). *A teaching handbook for university faculty. Chapter 1: Course objectives.* Retrieved October 1, 1998 from Indiana University Purdue University Indianapolis Web site: http://www.iupui.edu/~profdev/handbook/chap1.html

Illustrative Action Verbs for Defining Objectives using Bloom's Taxonomy

Taxonomy Categories	Sample Verbs for Stating Specific Learning Outcomes
Knowledge	Cite, define, identify, label, list, match, name, recognize, reproduce, select, state
Comprehension	Classify, convert, describe, distinguish between, explain, extend, give examples, illustrate, interpret, paraphrase, summarize, translate
Application	Apply, arrange, compute, construct, demonstrate, discover, modify, operate, predict, prepare, produce, relate, show, solve, use
Analysis	Analyze, associate, determine, diagram, differentiate, discriminate, distinguish, estimate, infer, order, outline, point out, separate, subdivide
Synthesis	Combine, compile, compose, construct, create, design, develop, devise, formulate, integrate, modify, organize, plan, propose, rearrange, reorganize, revise, rewrite, tell, write
Evaluation	Appraise, assess, compare, conclude, contrast, criticize, discriminate, evaluate, judge, justify, support, weigh

Reference:

Gronlund, N. E. (1998). *Assessment of student achievement.* Boston: Allyn and Bacon.

Examples of Instructional Objectives for the Cognitive Domain

1. The student will recall the four major food groups without error.
(Knowledge)

2. By the end of the semester, the student will summarize the main events of a story in grammatically correct English.
(Comprehension)

3. Given a presidential speech, the student will be able to point out the positions that attack a political opponent personally rather than the opponent's political programs.
(Analysis)

4. Given a short story, the student will write a different but plausible ending.
(Synthesis)

5. Given fractions not covered in class, the student will multiply them on paper with 85 percent accuracy.
(Application)

6. Given a description of a country's economic system, the student will defend it by basing arguments on principles of socialism.
(Evaluation)

7. From memory, with 80 percent accuracy the student will match each United States General with his most famous battle.
(Knowledge)

8. The student will describe the interrelationships among acts in a play.
(Analysis)

Reference:

Kubiszyn, K., & Borich, G. (1984). *Educational testing and measurement: Classroom application and practice.* Glenview, IL: Scott, Foresman, pp. 53-55.

Resources on Bloom's Taxonomy of the Cognitive Domain and Writing Educational Objectives

Bloom, B.S. (1956). *Taxonomy of educational objectives, Vol. 1*. New York: McKay.

Jacobs, L.C. and Chase, C.I. (1992). *Developing and using tests effectively: A guide for faculty*. San Francisco: Jossey-Bass.

Web resources:

Allen, T. (1998). *The taxonomy of educational objectives.* Retrieved November 3, 2003 from the Humbolt State University Web site: http://www.humboldt.edu/~tha1/bloomtax.html

Bixler, B. (2002). *Writing educational goals and objectives.* Retrieved November 3, 2003 from the Pennsylvania State University Web site: http://www.personal.psu.edu/staff/b/x/bxb11/Objectives/

Bloom's taxonomy. (2003). Retrieved November 3, 2003 from the University of Victoria Counseling Services Web site: http://www.coun.uvic.ca/learn/program/hndouts/bloom.html

Clark, D. (2002). *Learning domains or Bloom's taxonomy.* Retrieved November 3, 2003 from http://www.nwlink.com/~donclark/hrd/bloom.html

Huitt, W. (2000). *Bloom et al.'s taxonomy of the cognitive domain.* Retrieved November 3, 2003 from Valdosta State University Educational Psychology Web site: http://chiron.valdosta.edu/whuitt/col/cogsys/bloom.html

Krumme, G. (2001). *Major categories on the taxonomy of educational objectives.* Retrieved November 3, 2003 from the University of Washington Web site: http://faculty.washington.edu/krumme/guides/bloom.html

Writing educational goals and objectives. (2001). Retrieved November 3, 2003 from the University of Mississippi School of Pharmacy Bureau of Pharmaceutical Services Web site: http://www.pharmd.org/thebureau/N.htm

Test Blueprint

Once you know the learning objectives and item types you want to include in your test you should create a test blueprint. A test blueprint, also known as test specifications, consists of a matrix, or chart, representing the number of questions you want in your test within each topic and level of objective. The blueprint identifies the objectives and skills that are to be tested and the relative weight on the test given to each. The blueprint can help you ensure that you are obtaining the desired coverage of topics and level of objective. Once you create your test blueprint you can begin writing your items!

Example: 40 item exam

	Topic A	Topic B	Topic C	Topic D	TOTAL
Knowledge	1	2	1	1	5 (12.5%)
Comprehension	2	1	2	2	7 (17.5%)
Application	4	4	3	4	15 (37.5%)
Analysis	3	2	3	2	10 (25%)
Synthesis		1		1	2 (5%)
Evaluation			1		1 (2.5%)
TOTAL	10 (25%)	10 (25%)	10 (25%)	10 (25%)	40

Once you create your blueprint you should write your items to match the level of objective within each topic area.

Description of Multiple-Choice Items

Multiple-Choice Items:

Multiple-choice items can be used to measure knowledge outcomes and various types of learning outcomes. They are most widely used for measuring knowledge, comprehension, and application outcomes.

The multiple-choice item provides the most useful format for measuring achievement at various levels of learning. When selection-type items are to be used (multiple-choice, true-false, matching, check all that apply) an effective procedure is to start each item as a multiple-choice item and switch to another item type only when the learning outcome and content make it desirable to do so. For example, (1) when there are only two possible alternatives, a shift can be made to a true-false item; and (2) when there are a number of similar factors to be related, a shift can be made to a matching item.

Strengths:
1. Learning outcomes from simple to complex can be measured.
2. Highly structured and clear tasks are provided.
3. A broad sample of achievement can be measured.
4. Incorrect alternatives provide diagnostic information.
5. Scores are less influenced by guessing than true-false items.
6. Scores are more reliable than subjectively scored items (e.g., essays).
7. Scoring is easy, objective, and reliable.
8. Item analysis can reveal how difficult each item was and how well it discriminated between the strong and weaker students in the class
9. Performance can be compared from class to class and year to year
10. Can cover a lot of material very efficiently (about one item per minute of testing time).
11. Items can be written so that students must discriminate among options that vary in degree of correctness.
12. Avoids the absolute judgments found in True-False tests.

Limitations:
1. Constructing good items is time consuming.
2. It is frequently difficult to find plausible distractors.
3. This item is ineffective for measuring some types of problem solving and the ability to organize and express ideas.
4. Real-world problem solving differs – a different process is involved in proposing a solution versus selecting a solution from a set of alternatives.
5. Scores can be influenced by reading ability.
6. There is a lack of feedback on individual thought processes – it is difficult to determine why individual students selected incorrect responses.
7. Students can sometimes read more into the question than was intended.
8. Often focus on testing factual information and fails to test higher levels of cognitive thinking.
9. Sometimes there is more than one defensible “correct” answer.

10. They place a high degree of dependence on the student's reading ability and the instructor's writing ability.
11. Does not provide a measure of writing ability.
12. May encourage guessing.

Helpful Hints:

- Base each item on an educational or instructional objective of the course, not trivial information.
- Try to write items in which there is one and only one correct or clearly best answer.
- The phrase that introduces the item (stem) should clearly state the problem.
- Test only a single idea in each item.
- Be sure wrong answer choices (distractors) are at least plausible.
- Incorporate common errors of students in distractors.
- The position of the correct answer should vary randomly from item to item.
- Include from three to five options for each item.
- Avoid overlapping alternatives (see Example 3 following).
- The length of the response options should be about the same within each item (preferably short).
- There should be no grammatical clues to the correct answer.
- Format the items vertically, not horizontally (i.e., list the choices vertically)
- The response options should be indented and in column form.
- Word the stem positively; avoid negative phrasing such as "not" or "except." If this cannot be avoided, the negative words should always be highlighted by underlining or capitalization: Which of the following is <u>NOT</u> an example ……
- Avoid excessive use of negatives and/or double negatives.
- Avoid the excessive use of "All of the above" and "None of the above" in the response alternatives. In the case of "All of the above", students only need to have partial information in order to answer the question. Students need to know that only two of the options are correct (in a four or more option question) to determine that "All of the above" is the correct answer choice. Conversely, students only need to eliminate one answer choice as implausible in order to eliminate "All of the above" as an answer choice. Similarly, with "None of the above", when used as the correct answer choice, information is gained about students' ability to detect incorrect answers. However, the item does not reveal if students know the correct answer to the question.

Example 1

The stem of the original item below fails to present the problem adequately or to set a frame of reference for responding.

Original

1. World War II was:

   A. The result of the failure of the League of Nations.
   B. Horrible.
   C. Fought in Europe, Asia, and Africa.
   D. Fought during the period of 1939-1945.

Revised

1. In which of these time period was World War II fought?

   A. 1914-1917
   B. 1929-1934
   C. 1939-1945
   D. 1951-1955
   E. 1961-1969

Example 2

There should be no grammatical clues to the correct answer.

Original

1. Albert Eisenstein was a:

   A. Anthropologist.
   B. Astronomer.
   C. Chemist.
   D. Mathematician

Revised

1. Who was Albert Einstein?

   A. An anthropologist.
   B. An Astronomer.
   C. A chemist.
   D. A mathematician.

Example 3

Alternatives should not overlap (e.g., in the original form of this item, if either of the first two alternatives is correct, “C” is also correct.)

Original

1. During what age period is thumb-sucking likely to produce the greatest psychological trauma?

   A. Infancy
   B. Preschool period
   C. Before adolescence
   D. During adolescence
   E. After adolescence

Revised

1. During what age period is thumb-sucking likely to produce the greatest psychological trauma?

   A. From birth to 2 years old
   B. From 2 years to 5 years old
   C. From 5 years to 12 years old
   D. From 12 years to 20 years old
   E. 20 years of age or older

Example 4

Example of how the greater similarity among alternatives increases the difficulty of the item.

Easy

1. Who was the President of the U.S. during the War of 1812?

A. Grover Cleveland
B. Abraham Lincoln
C. James Madison
D. Harry Truman
E. George Washington

More Difficult

1. Who was President of the U.S. during the War of 1812?

A. John Q. Adams
B. Andrew Jackson
C. Thomas Jefferson
D. James Madison
E. George Washington

Reference:

Marshall, J. C., & Hales, L. W. (1971). *Classroom test construction*. Reading MA: Addison-Wesley.

Multiple-Choice Item Writing Guidelines

Multiple-choice questions typically have 3 parts: a stem, the correct answer – called the key, and several wrong answers, called distractors.

Procedural Rules:

- Use either the best answer or the correct answer format.
  - Best answer format refers to a list of options that can all be correct in the sense that each has an advantage, but one of them is the best.
  - Correct answer format refers to one and only one right answer.
- Format the items vertically, not horizontally (i.e., list the choices vertically)
- Allow time for editing and other types of item revisions.
- Use good grammar, punctuation, and spelling consistently.
- Minimize the time required to read each item.
- Avoid trick items.
- Use the active voice.
- The ideal question will be answered by 60-65% of the tested population.
- Have your questions peer-reviewed.
- Avoid giving unintended cues – such as making the correct answer longer in length than the distractors.

Content-related Rules:

- Base each item on an educational or instructional objective of the course, not trivial information.
- Test for important or significant information.
- Focus on a single problem or idea for each test item.
- Keep the vocabulary consistent with the examinees' level of understanding.
- Avoid cueing one item with another; keep items independent of one another.
- Use the author's examples as a basis for developing your items.
- Avoid overly specific knowledge when developing items.
- Avoid textbook, verbatim phrasing when developing the items.
- Avoid items based on opinions.
- Use multiple-choice to measure higher level thinking.
- Be sensitive to cultural and gender issues.
- Use case-based questions that use a common text to which a set of questions refers.

Stem Construction Rules:

- State the stem in either question form or completion form.
- When using a completion form, don't leave a blank for completion in the beginning or middle of the stem.

- Ensure that the directions in the stem are clear, and that wording lets the examinee know exactly what is being asked.
- Avoid window dressing (excessive verbiage) in the stem.
- Word the stem positively; avoid negative phrasing such as "not" or "except." If this cannot be avoided, the negative words should always be highlighted by underlining or capitalization: Which of the following is NOT an example ……
- Include the central idea and most of the phrasing in the stem.
- Avoid giving clues such as linking the stem to the answer (…. Is an example of *an*: test-wise students will know the correct answer should start with a vowel)

General Option Development Rules:

- Place options in logical or numerical order.
- Use letters in front of options rather than numbers; numerical answers in numbered items may be confusing to students.
- Keep options independent; options should not be overlapping.
- Keep all options homogeneous in content.
- Keep the length of options fairly consistent.
- Avoid, or use sparingly, the phrase *all of the above.*
- Avoid, or use sparingly, the phrase *none of the above*.
- Avoid the use of the phrase *I don't know*.
- Phrase options positively, not negatively.
- Avoid distractors that can clue test-wise examinees; for example, absurd options, formal prompts, or semantic (overly specific or overly general) clues.
- Avoid giving clues through the use of faulty grammatical construction.
- Avoid specific determinates, such as *never* and *always*.
- Position the correct option so that it appears about the same number of times in each possible position for a set of items.
- Make sure that there is one and only one correct option.

Distractor (incorrect options) Development Rules:

- Use plausible distractors.
- Incorporate common errors of students in distractors.
- Avoid technically phrased distractors.
- Use familiar yet incorrect phrases as distractors.
- Use true statements that do not correctly answer the item.
- Avoid the use of humor when developing options.
- Distractors that are not chosen by any examinees should be replaced.

Suggestions for Writing Good Multiple Choice Items:

- Present practical or real-world situations to the students.
- Present the student with a diagram of equipment and ask for application, analysis or evaluation.

- Present actual quotations taken from newspapers or other published sources and ask for the interpretation or evaluation of these quotations.
- Use pictorial materials that require students to apply principles and concepts.
- Use charts, tables or figures that require interpretation.

References:

Carneson, J., Delpierre, G., & Masters, K. (n.d.). *Designing and managing multiple choice questions: Appendix B, designing MCQs – do's and don'ts.* Retrieved November 3, 2003 from the University of Cape Town Web site: http://www.uct.ac.za/projects/cbe/mcqman/mcqappb.html

Haladyna, T. M. (1999). *Developing and validating multiple-choice test items, 2nd ed*. Mahwah, NJ: Lawrence Erlbaum Associates.

Haladyna, T. M. (1989). Taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2 (1),* 37-50.

Jacobs, L. C. (2002). *How to write better tests: A handbook for improving test construction skills.* Retrieved November 3, 2003 from Indiana University Bloomington Evaluation Services & Testing Web site: http://www.indiana.edu/~best/write_better_tests.shtml

Sevenair. J. P., & Burkett, A. R. (1997). *Item writing guidelines.* Retrieved November 3, 2003 from the Xaviar University of Louisiana Web site: http://webusers.xula.edu/jsevenai/objective/guidelines.html

Writing multiple choice questions that demand critical thinking. (2002). Retrieved November 3, 2003 from the University of Oregon Teaching Effectiveness Program Web site: http://tep.uoregon.edu/resources/assessment/multiplechoicequestions/mc4critthink.html

Guidelines to Writing Test Items

- Begin writing items well ahead of the time when they will be used; allow time for revision.
- Match items to intended outcomes at the proper difficulty level to provide a valid measure of the instructional objectives.
- Be sure each item deals with an important aspect of the content area and not with trivia.
- Be sure that the problem posed is clear and unambiguous.
- Be sure that each item is independent of all other items (i.e., a hint to an answer should not be unintentionally embedded in another item).
- Be sure the item has one correct or best answer on which experts would agree.
- Prevent unintended clues to the answer in the statement or question (e.g., grammatical inconsistencies such as ‘a’ or ‘an’ give clues).
- Avoid duplication of the textbook in writing test items; don’t lift quotes directly from any textual materials.
- Avoid trick or catch questions in an achievement test. (Don’t waste time testing how well the student can interpret your intentions).
- On a test with different question formats (e.g., multiple choice and True-False), one should group all items of similar format together.
- Questions should follow an easy to difficult progression.
- Space the items to eliminate overcrowding.
- Have diagrams and tables <u>above</u> the item using the information, <u>not</u> below.

Preparing Your Students for Taking Multiple-Choice Tests

1. Specify objectives or give study questions
   - Students should not be forced to guess what will be on a test
   - Give students specific study questions or topics, then draw the test items from those questions
   - There should be many study questions that are comprehensive and cover all the important ideas in the course

2. Try to reduce frustration for the creative student
   - Avoid "all of these," "none of these," and "both A and B" answer choices.

3. Defeat the "test-wise" strategies of students who don't study
   - Be aware of the general "rules of thumb" that students use to guess on multiple choice exams and try to avoid them

   a. Pick the longest answer
      - make sure the longest answer is only correct a part of the time
      - try to make options equal length

   b. When in doubt pick "c"
      - make sure the correct answer choice letter varies

   c. Never pick an answer which uses the word 'always' or 'never' in it
      - make sure this option is correct part of the time or avoid using always and never in the option choices

   d. If there are two answers which express opposites, pick one or the other and ignore other alternatives
      - sometimes offer opposites when neither is correct or offer two pairs of opposites

   e. If in doubt, guess
      - use five alternatives instead of three or four to reduce guessing

   f. Pick the scientific-sounding answer
      - use scientific sounding jargon in wrong answers

   g. Don't pick an answer which is too simple or obvious
      - sometimes make the simple, obvious answer the correct one

   h. Pick a word which you remember was related to the topic
      - when creating the distractors use terminology from the same area of the text as the right answer, but in distractors use those words incorrectly so the wrong answers are definitely wrong

Reference:

Dewey, R. A. (1998, January 20). *Writing multiple choice items which require comprehension.* Retrieved November 3, 2003 from http://www.psywww.com/selfquiz/aboutq.htm

Sample Multiple-Choice Items Related to Bloom's Taxonomy

Knowledge Items:

Outcome: Identifies the meaning of a term.

Reliability is the same as:

A. consistency.
B. relevancy.
C. representativeness.
D. usefulness.

Outcome: Identifies the order of events.

What is the first step in constructing an achievement test?

A. Decide on test length.
B. Identify the intended learning outcomes.
C. Prepare a table of specifications.
D. Select the item types to use.

Comprehension Items:

Outcome: Identifies an example of a term.

Which one of the following statements contains a specific determiner?

A. America is a continent.
B. America was discovered in 1492.
C. America has some big industries.
D. America's population is increasing.

Outcome: Interprets the meaning of an idea.

The statement that "test reliability is a necessary but not sufficient condition of test validity" means that:

A. a reliable test will have a certain degree of validity.
B. a valid test will have a certain degree of reliability.
C. a reliable test may be completely invalid and a valid test completely unreliable.

Outcome: Identifies an example of a concept or principle.

Which of the following is an example of a criterion-referenced interpretation?

A. Derik earned the highest score in science.
B. Erik completed his experiment faster than his classmates.
C. Edna's test score was higher than 50 percent of the class.
D. Tricia set up her laboratory equipment in five minutes.

Outcome: Predicts the most probable effect of an action.

What is most likely to happen to the reliability of the scores for a multiple-choice test, where the number of alternatives for each item is changed from three to four?

A. It will decrease.
B. It will increase.
C. It will stay the same.
D. There is no basis for making a prediction.

Application Items

Outcome: Distinguishes between properly and improperly stated outcomes.

Which one of the following learning outcomes is properly stated in terms of student performance?

A. Develops an appreciation of the importance of testing.
B. Explains the purpose of test specifications.
C. Learns how to write good test items.
D. Realizes the importance of validity.

Outcome: Improves defective test items.

Directions: read the following test item and then indicate the best change to make to improve the item.

Which one of the following types of learning outcomes is most difficult to evaluate objectively?

1. A concept.
2. An application.
3. An appreciation.
4. None of the above.

The best change to make in the previous item would be to:

A. change the stem to incomplete-statement form.
B. use letters instead of numbers for each alternative.
C. remove the indefinite articles "a" and "an" from the alternatives.
D. replace "none of the above" with "an interpretation."

Analysis Items

Directions: Read the following comments a teacher made about testing. Then answer the questions that follow by circling the letter of the best answer.

"Students go to school to learn, not to take tests. In addition, tests cannot be used to indicate a student's absolute level of learning. All tests can do is rank students in order of achievement, and this relative ranking is influenced by guessing, bluffing, and the subjective opinions of the teacher doing the scoring. The teacher-learning process would benefit if we did away with tests and depended on student self-evaluation."

Outcome: Recognizes unstated assumptions.

1. Which one of the following unstated assumptions is this teacher making?

A. Students go to school to learn.
B. Teachers use essay tests primarily.
C. Tests make no contribution to learning.
D. Tests do not indicate a student's absolute level of learning.

Outcome: Identifies the meaning of a term.

2. Which one of the following types of test is this teacher primarily talking about?

A. Diagnostic test.
B. Formative test.
C. Pretest.
D. Summative test.

Synthesis Item

(See paragraph for analysis items)

Outcome: Identifies relationships.

3. Which one of the following propositions is most essential to the final conclusion?

A. Effective self-evaluation does not require the use of tests.
B. Tests place students in rank order only.
C. Test scores are influenced by factors other than achievement.
D. Students do not go to school to take tests.

Reference:

Gronlund, N. E. (1998). *Assessment of student achievement*. Boston: Allyn and Bacon.

More Sample Multiple-Choice Items

Knowledge:

1. In the area of physical science, which one of the following definitions describes the term "polarization"?

A. The separation of electric charges by friction.
B. The ionization of atoms by high temperatures.
C. The interference of sound waves in a closed chamber.
D. The excitation of electrons by high frequency light.
E. The vibration of transverse waves in a single plane.

Simple recall of the correct definition of polarization is required.

Comprehension:

2. Which one of the following describes what takes place in the so-called PREPARATION stage of the creative process, as applied to the solution of a particular problem?

A. The problem is identified and defined.
B. All available information about the problem is collected.
C. An attempt is made to see if the proposed solution to the problem is acceptable.
D. The person goes through some experience leading to a general idea of how the problem can be solved.
E. The person sets the problem aside, and gets involved with some other unrelated activity.

The knowledge of the five stages of the creative process must be recalled (knowledge) and the student is tested for an understanding (comprehension) of the meaning of each term, in this case, "preparation."

Application:

3. Which one of the following memory systems does a piano-tuner mainly use in his occupation?

A. Echoic memory
B. Short-term memory
C. Long-term memory
D. Mono-auditory memory
E. None of the above

This question tests for the application of previously acquired knowledge (the various memory systems).

Analysis:

4. Read carefully through the paragraph below, and decide which of the options A-D is correct.

"The basic premise of pragmatism is that questions posed by speculative metaphysical propositions can often be answered by determining what the practical consequences of the acceptance of a particular metaphysical proposition are in this life. Practical consequences are taken as the criterion for assessing the relevance of all statements or ideas about truth, norm and hope."

A. The word "acceptance" should be replaced by "rejection."
B. The word "often" should be replaced by "only."
C. The word "speculative" should be replaced by hypothetical."
D. The word "criterion" should be replaced by "measure."

This question requires prior knowledge of and understanding about the concept of pragmatism. The student is tested on his/her ability to analyze whether a word fits with the accepted definition of pragmatism.

Evaluation:

5. Judge the sentence in italics according to the criteria given below:

"The United States took part in the Gulf War against Iraq BECAUSE of the lack of civil liberties imposed on the Kurds by Saddam Hussein's regime."

A. The assertion and the reason are both correct, and the reason is valid.
B. The assertion and the reason are both correct, but the reason is invalid.
C. The assertion is correct but the reason is incorrect.
D. The assertion is incorrect but the reason is correct.
E. Both the assertion and the reason are incorrect.

A knowledge and understanding of Middle East politics is assumed. The student is tested in the ability to evaluate between cause and effect in the sentence in terms of predefined criteria.

Reference:

Carneson, J., Delpierre, G., & Masters, K. (n.d.). *Designing and managing multiple choice questions: Appendix C, multiple choice questions and Bloom's taxonomy*. Retrieved November 3, 2003 from the University of Cape Town Web site: http://www.uct.ac.za/projects/cbe/mcqman/mcqappc.html

Good versus Poor Multiple-Choice Items

Presented here are some possible multiple-choice questions in areas of statistics, biology and communication. Even though each question within the pair assesses the same content, those in bold encourage the learner to think about the problem in more depth, to use and apply their knowledge and not simply recall memorized information. Because the questions in bold encourage more meaningful processing of information they are considered more effective. *The correct answers have been starred.

Statistics

1a. The mean of a distribution of test scores is the:

a) Most frequently occurring score
b) 50^{th} percentile
c) Arithmetic average*
d) Measure of score range

1b. A university developed an aptitude test to use for admission to its Honors Program. The test was administered to a group of seven applicants who obtained the following scores: 70,72,72,80,89,94,98. The mean score on the aptitude test is:

a) 72
b) 80
c) 82*
d) 90

Biology

1a. Suppose you thoroughly and adequately examined a particular type of cell, using the transmission electronic microscope, and discovered that it completely lacked ribosomes. You would then conclude that this cell also lacked:

a) A nucleus
b) DNA
c) Cellulose
d) Protein synthesis*

1b. Ribosomes are important for:

a) the nucleus
b) DNA
c) Cellulose
d) Protein synthesis*

Communication

1a. What is an example of pseudolistening (a pitfall to effective listening)?

a) daydreaming while nodding your head*
b) sidetracking a conversation
c) premature replying
d) paying attention to the context

1b. While Amy is presenting her proposal to the group, Josh is thinking about his weekend fishing trip. Even though he is not listening to a word Amy is saying, he manages to occasionally nod his head in agreement. Josh's behavior is an example of:

a) pseudolistening*
b) premature replying
c) attentiveness to the context
d) conversation sidetracking

Activity: Identifying Flawed Multiple-Choice Items

For each pair of items, decide which item is better and why.

1A. The promiscuous use of sprays, oils, and antiseptics in the nose during acute colds is a pernicious practice because it may have a deleterious effect on:

a. the sinuses
b. red blood cells
c. white blood cells
d. the olfactory nerve

1B. Frequent use of sprays, oils, and antiseptics in the nose during a bad cold may result in:

a. the spreading of the infection to the sinuses
b. damage to the olfactory nerve
c. destruction of white blood cells
d. congestion of the mucous membrane in the nose

2A. In 1965, the death rate from accidents of all types per 100,000 population in the 15-24 age group was:

a. 59.0
b. 59.1
c. 59.2
d. 59.3

2B. In 1965, the leading cause of death per 100,000 population in the 15-24 age group was from:

a. respiratory disease
b. cancer
c. accidents
d. rheumatic heart disease

3A. About how many calories are recommended daily for a 14-year old who is 62 in. tall, weighs 103 lbs., and is moderately active?

a. 1,500
b. 2,000
c. 2,500
d. 3,000

3B. About how many calories are recommended daily for a 14-year old who is 62 in. tall, weighs 103 lbs., and is moderately active?

a. 0
b. 2,000
c. 2,500
d. 3,000

4A. Which of the following is a category in the taxonomy of the cognitive domain?
A. Reasoning ability
B. Critical thinking
C. Rote learning
D. All of the above
E. None of the above

4B. What is the most complex level in the taxonomy of the cognitive domain?
A. Knowledge
B. Synthesis
C. Evaluation
D. Analysis
E. Comprehension

Answers: Identifying Flawed Multiple-Choice Items

Pair #1: 1B is the better item. 1A is wordy and uses vocabulary that may be unfamiliar to many students. 1B not only asks what part of the body is affected (sinuses) but also what is the result (spreading of infection).

Pair #2: 2B is the better item. The difference between the options in 2A is trivial. Also, 2A asks for memorization of factual information.

Pair #3: 3A is the better item. 3B contains a distractor that is not plausible (0 calories).

Pair #4: 4B is the better item. 4A asks for simple identification of a category, whereas 4B asks for differentiation between the levels. 4A also contains "all of the above" and "none of the above" as option choices, which should be avoided, if possible.

Pair #5: 5A is the better item. 5A sites the source of the information, whereas 5B can be construed as an opinion question.

Scenario-Based Problem Solving Item Set

Presented here are some scenario–based problem solving item sets in statistics and biology. This method provides a basis for testing complex thinking, application of knowledge as well as integration of material. It is a structured and well-organized method of assessment, ensuring ease of scoring. Note that it may be time consuming to write appropriate scenarios.

Statistics

Two researchers were studying the relationship between amount of sleep each night and calories burned on an exercise bike for 42 men and women. They were interested if people who slept more had more energy to use during their exercise session. They obtained a correlation of .28, which has a two-tailed probability of .08. Alpha was .10.

1. Which is an example of a properly written research question?

a) Is there a relationship between amount of sleep and energy expanded?*
b) Does amount of sleep correlate with energy used?
c) What is the cause of energy expanded?
d) What is the value of rho?

2. What is the correct term for the variable amount of sleep?

a) Dependent*
b) Independent
c) Predictor
d) y

3. What is the correct statistical null hypothesis?

a) There is no correlation between sleep and energy expanded
b) Rho equals zero*
c) R equals zero
d) Rho equals r

4. What conclusions should you draw regarding the null hypothesis?

a) Reject*
b) Accept
c) Cannot determine without more information

5. What conclusions should you draw regarding this study?

a) The correlation was significant
b) The correlation was not significant
c) A small relationship exists*
d) No relationship exists

Reference:

Haladyna, T. M. (1994). *Developing and validating-multiple choice test items, 1st ed*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Biology

One day you meet a student watching a wasp drag a paralyzed grasshopper down a small hole in the ground. When asked what he is doing he replies, "I'm watching that wasp store paralyzed grasshoppers in her nest to feed her offspring."

1. Which of the following is the best description of his reply?

a) He is not a careful observer.
b) He is stating a conclusion only partly derived from his observation.*
c) He is stating a conclusion entirely drawn from his observation.
d) He is making no assumptions.

2. Which of the following additional observations would add the most strength to the student's reply in Question 1?

a) Observing the wasp digging a similar hole.
b) Observing the wasp dragging more grasshoppers into the hole.
c) Digging into the hole and observing wasp eggs on the paralyzed grasshopper*
d) Observing adult wasps emerging from the hole a month later.

3. Both of you wait until the wasp leaves the area, then you dig into the hole and observe three paralyzed grasshoppers, each with a white egg on its side. The student states that this evidence supports his reply in Question 1. Which of the following assumptions is he making?

a) The eggs are grasshopper eggs.
b) The wasp laid the eggs.*
c) The wasp dug the hole.
d) The wasp will return with another grasshopper.

4. You take the white eggs to the Biology laboratory. Ten days later immature wasps hatched from the eggs. The student states that this evidence supports his reply in Question 1. Which of the following assumptions is he making?

a) The wasp dug the hole.
b) The wasp stung the grasshoppers.
c) The grasshoppers were dead.
d) A paralyzed grasshopper cannot lay an egg.*

Reference:

Donovan, M. P., & Allen, R.D. (n.d.). *Analytical problems in biology*. Morgantown, West Virginia: Alpha Editions.

Additional reading:

Terry, T.M. (1980). The narrative exam – an approach to creative organization of multiple-choice tests. *Journal of College Science Teaching*, *9(3)*, 156-158.

An Alternative Multiple-Choice Method

There are a number of different ways multiple choice exams can be used in the classroom. One such example comes from Stanford University. Presented here is a summary of the Human Biology Project implemented at this institution.

The Human Biology Project

The Stanford Learning Lab implemented a new approach to assess student learning by using weekly on-line problem sets for a Stanford Human Biology course, The Human Organism. The web-based problem sets created by the Stanford Learning Lab allowed a large lecture class to focus on the individual student, permitting personal and rapid feedback.

The Human Biology class at Stanford University is a large undergraduate course with 2 professors, 5 assistants and 208 students. It covers the topic of physiology, or how animals (including humans) work. The course consists of four one-hour lectures and one-hour discussion section each week.

During Spring Quarter 1998, the faculty team provided a problem set to the students via the Web at the end of each week's lecture. Graded by computer, the correct answers to the sets were posted on the Web. In addition to selecting a multiple-choice answer to each question, students were required to submit a short "rationale" explaining their answers. The faculty team sorted responses to make it easier to explore frequently-missed questions. The course assistant then used this information to tailor class instruction.

Sample Question and Rationale:

1. Which of the following has/have intrinsic pacemaker characteristics?

a) Medulla c) Sinoatrial node
b) Pons d) Atrioventricular node

Ideal rationale:

SA node is the normal pacemaker for the entire heart. AV node also has pacemaker potential, but is overshadowed by SA node. Medulla has pacemaker potential for breathing rhythm as well. Pons helps refine rhythm, but does not have pacemaker potential.

Less–than-ideal rationales:

Offering an incomplete answer:
Normally the SA node is responsible for generating heart rate, and it is able to do this because of its intrinsic rhythm. The AV node also has an intrinsic rhythm, but it is "overshadowed" by that of the SA node.

Providing a quotation from the book:
The sinoatrial node is the pacemaker of the mammalian heart.

Providing irrelevant information:
Stretch receptors are located in the aortic arch and the carotid sinus. They have the ability to respond to changes in pressure.

Restating the answer:
The SA node, AV node, and medulla all possess intrinsic pacemaker characteristics as they all serve as intrinsic pacemakers.

Blind appeal to authority:
This answer is right because Professor Heller said that it was, and Professor Heller is cool.

Worst rationale:
No rationale submitted

References:

Schaeffer, E., Michalchik, V., Martin, M. Birks, H., & Nash, J. (1999). *Web-based problem sets in the human biology program: Fall 1998*. Retrieved November 3, 2003 from Stanford University, Stanford Learning Lab Web site: http://sll.stanford.edu/projects/humbioa/HumBio_2a2b_98.pdf

Nash, J. & Schaeffer, E. (1999, January 11). Web-based coursework proves useful. *Speaking of Computers*, *49*. Retrieved November 3, 2003 from http://acomp.stanford.edu/acpubs/SOC/Back_Issues/SOC49/humbio.html

Guidelines for Administering Examinations

One of the basic problems in education is determining how well the students have learned the material covered in the course. It is quite possible that a particular student may know the material being tested extremely well but still perform poorly on examinations. If one conceives of an examination as a measurement device analogous to, for example, a ruler, then the accuracy of the assessment of how "well" someone knows the course material is a function of the quality of the examination. However, the actual administration of the examination may also affect a student's performance. Presented below is a list of general principles to consider when designing and administering examinations.

1. Give complete instructions as to how the examination is to be taken. It is helpful to students for you to indicate the number of points each section of the examination counts or the amount of time to spend on each question. This indication of the relative importance of each question helps them to allocate their time and efforts wisely.

2. If the student is allowed to take any aids into the examination room (such as a calculator, notes, textbook), be sure to state <u>specifically</u> what is allowed.

3. The examination should test the lesson or course objectives. The lesson assignments themselves should provide preparation for taking the final examination in content as well as in the practice of answering certain kinds of questions. For example, if the lesson assignments ask all essay questions, it would be inappropriate for the examination to consist of 200 multiple-choice questions. Practice taking the completed test yourself. You should count on the students to take about four times the amount of time it takes you to complete the test.

4. For final examinations, structure the test to cover the scope of the entire course. The examination should be comprehensive enough to test adequately the student's learning of the course material. It is usually a good idea to use a variety of different types of questions on the examination (e.g., multiple-choice, essay, etc.) because certain subject matter areas can be covered most effectively with certain types of items. However, items of the same type should be kept together when possible.

5. Prior to the examination, tell the students what types of questions will be on the test (essay, multiple-choice, etc.). If possible, it is a good practice to allow students access to past (retied) examinations so that they have some idea what to expect. Also, if you plan to administer essay exams, it is a good idea to share the general grading scheme ahead of time so that the students know the criteria by which they will be evaluated.

6. Present to the students a list of review questions or a list of topics to be covered on the examination along with an indication of the relative emphasis on each.

7. Give detailed study suggestions.

8. Indicate how much the examination will count toward determining the final grade.

Analyzing Multiple-Choice Item Responses

Understanding how to interpret and use information based on student test scores is as important as knowing how to construct a well-designed test. Using feedback from your test to guide and improve instruction is an essential part of the process.

Using statistical information to review your multiple-choice test can provide useful information. Three of these statistics are:

Item difficulty: the percentage of students that correctly answered the item.

- Also referred to as the p-value.
- The range is from 0% to 100%, or more typically written as a proportion as 0.0 to 1.00.
- The higher the value, the *easier* the item.
- Calculation: Divide the number of students who got an item correct by the total number of students who answered it.
- P-values above 0.90 are very easy items and should not be reused again for subsequent tests. If almost all of the students can get the item correct, it is a concept probably not worth testing.
- P-values below 0.20 are very difficult items and should be reviewed for possible confusing language, removed from subsequent tests, and/or highlighted for an area for re-instruction. If almost all of the students get the item wrong there is either a problem with the item or students did not get the concept.
- Ideal value: Slightly higher than midway between chance (1.00 divided by the number of choices) and a perfect score (1.00) for the item. For a 5-option multiple-choice question the ideal value is .60 (60%)

Item discrimination: the point-biserial relationship between how well students did on the item and their total test score.

- Also referred to as the Point-Biserial correlation (PBS)
- The range is from –1.00 to 1.00.
- The higher the value, the more discriminating the item. A highly discriminating item indicates that the students who had high tests scores got the item correct whereas students who had low test scores got the item incorrect.
- Items with discrimination values near or less than zero should be removed from the test. This indicates that students who overall did poorly on the test did *better* on that item than students who overall did well. The item may be confusing for your better scoring students in some way.
- Acceptable range: 0.20 or higher
- Ideal value: The closer to 1.00 the better
- Calculation: $\frac{(\overline{X}_C - \overline{X}_T)}{S.D.Total}\sqrt{\frac{p}{q}}$ where

  $\overline{X}_C$ = the mean total score for persons who have responded correctly to the item

  $\overline{X}_T$ = the mean total score for all persons

p = the difficulty value for the item
$q = (1 - p)$
S. D. Total = the standard deviation of total test scores

Reliability coefficient: a measure of the amount of measurement error associated with a test score.

- The range is from 0.0 to 1.0.
- The higher the value, the more reliable the overall test score.
- Typically, the internal consistency reliability is measured. This indicates how well the items are correlated with one another.
- High reliability indicates that the items are all measuring the same thing, or general construct (e.g. knowledge of how to calculate integrals for a Calculus course).
- With multiple-choice items that are scored correct/incorrect, the Kuder-Richardson formula 20 (KR-20) is often used to calculate the internal consistency reliability.
    - $\frac{K}{K-1}(1-\frac{\sum pq}{\sigma^2_x})$ where

      K = number of items
      p = proportion of persons who responded correctly to an item (i.e., difficulty value)
      q = proportion of persons who responded incorrectly to an item (i.e., 1 – p)
      σ^2_x = total score variance
- Two ways to improve the reliability of the test are to 1) increase the number of questions in the test or 2) use items that have high discrimination values in the test
- Acceptable range: 0.60 or higher
- Ideal value: 1.00

Another useful item review technique to use is **distractor evaluation**.

The distractor should be considered an important part of the item. Nearly 50 years of research shows that there is a relationship between the distractors students choose and total test score. The quality of the distractors influence student performance on a test item. Although the correct answer must be truly correct, it is just as important that the distractors be incorrect. Distractors should appeal to low scorers who have not mastered the material whereas high scorers should infrequently select the distractors. Reviewing the options can reveal potential errors of judgment and inadequate performance of distractors. These poor distractors can be revised, replaced, or removed.

One way to study responses to distractors is with a frequency table. This table tells you the number and/or percent of students that selected a given distractor. Distractors that are selected by a few or no students should be removed or replaced. These kinds of distractors are likely to be so implausible to students that hardly anyone selects them.

- Definition: The incorrect alternatives in a multiple-choice item.
- Reported as: The frequency (count), or number of students, that selected each incorrect alternative
- Acceptable Range: Each distractor should be selected by at least a few students
- Ideal Value: Distractors should be equally popular
- Interpretation:
    - Distractors that are selected by a few or no students should be removed or replaced
    - One distractor that is selected by as many or more students than the correct answer may indicate a confusing item and/or options
- The number of people choosing a distractor can be lower or higher than the expected because:
    - Partial knowledge
    - Poorly constructed item
    - Distractor is outside of the area being tested

References:

DeVellis, R. F. (1991). *Scale development: Theory and applications*. Newbury Park: Sage Publications.

Haladyna. T. M. (1999). *Developing and validating multiple-choice test items, 2nd ed*. Mahwah, NJ: Lawrence Erlbaum Associates.

Suen, H. K. (1990). *Principles of test theories*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Activity: Item Analysis

Below is a sample item analysis performed by MEC that shows the summary table of item statistics for all items for a multiple-choice classroom exam. Review the item difficulty (P), discrimination (R(IT)), and distractors (options B-E).

Item Analysis (sample of 10 items) – correct answer is "A"

Summary Table of Test Item Statistics

<test name>

N TOTAL = 932 MEAN TOTAL = 69.4 S.D. TOTAL = 10.2 ALPHA = .84

ITEM	P	R(IT)	NC	MC	MI	OMIT	A	B	C	D	E
1.	0.72	0.34	667	71.56	67.66	1	**667**	187	37	30	10
2.	0.90	0.21	840	70.11	69.02	1	**840**	1	76	9	5
3.	0.60	0.39	561	72.66	65.47	0	**561**	233	46	88	4
4.	0.99	-0.06	923	69.34	69.90	0	**923**	3	3	3	0
5.	0.94	0.14	876	69.76	68.23	0	**876**	0	12	24	20
6.	0.77	-0.01	716	69.34	69.57	0	**716**	16	25	35	140
7.	0.47	0.31	432	72.76	66.16	3	**432**	107	68	165	157
8.	0.12	0.08	114	71.61	68.39	8	**114**	218	264	153	175
9.	0.08	0.04	75	70.78	69.03	0	**75**	64	120	67	606
10.	0.35	0.42	330	75.24	63.54	0	**330**	98	74	183	247
.											
.											
40.											

Which item(s) would you remove altogether from the test? Why?

Which distractor(s) would you revise? Why?

Which items are working well?