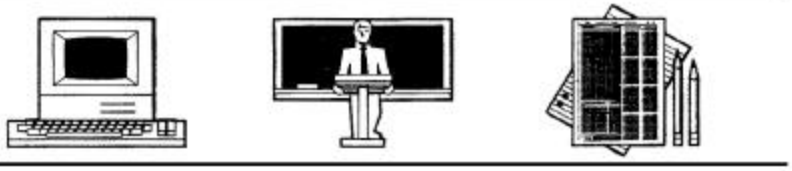
HANDBOOK ON TESTING AND GRADING



Office of Instructional Resources, University of Florida, Gainesville, Florida 32611

OIR

Preface

When we test we know which students earned higher scores than other students. Without a well constructed examination, however, we may not know what students did or did not learn, nor do we necessarily know to what degree the test score is a reasonably accurate estimate of student achievement. If testing is to be related to learning in a meaningful way, the entire testing process needs to be carefully integrated into the organization of the course. Hopefully, good testing will (1) promote student learning, (2) be used by instructors to improve their courses, and (3) result in grades that are a valid indicator of how well students have learned the course material. The assignment of grades then takes third priority as a reason for giving an examination.

Student assessment can take many forms. This handbook focuses on how to construct and assess machine-scored multiple choice items, short answer items and essay questions. Techniques for rating students on oral examinations or other performance measures, while important, are beyond the present scope of the handbook. Of specific concern in the handbook are techniques for writing and evaluating good test questions in a variety of formats that assess the student's ability not only to recall facts, but also to analyze and solve problems.

In addition to helping instructors construct tests and assess student achievement more accurately, the handbook includes chapters on procedures for successfully administering tests, analyzing test questions, and assigning grades.

Sue M. Legg, Ph.D.

January, 1991

Table of Contents

I. Steps In Writing Examinations

■ *Developing the Test Blueprint*

The first step in test development is to set the test specifications based on the relative importance of the content to be tested. The usual procedure is to develop a test blueprint which includes the test objectives and the cognitive level of the items. The test objectives are weighted by assigning a percentage of the test items to each objective. Thus, a test that covers five areas equally would have twenty percent of the items assigned to each objective. Some objectives may emphasize factual knowledge while others stress understanding or application of knowledge. Therefore, it is useful to place the objectives on one axis of the blueprint and the cognitive level on the other axis. In this way the test can be balanced by content and cognitive requirements.

At this point, the instructor should review the length of the planned examination to be certain students can complete it in the time allowed. While speed in taking examinations could be relevant in some subject areas, speeded tests discriminate against the high ability but more methodical student. As a rule of thumb, students can respond to one relatively complex multiple choice item every 50 seconds. Items requiring calculation may take longer. Time for writing responses to an essay question also depend on the complexity of the task. An instructor might double the time for the class that it takes for the instructor to write an acceptable response.

In the sample blueprint in Table 1 below, four content areas and three cognitive levels are to be tested. Eighteen percent of the 50 items (9 items) relate to the objective of patient comfort. Thirty percent of the test (15 items) require a knowledge of terminology in the four content areas.

Table 1

DEVELOP A TEST BLUEPRINT				
	COGNITIVE LEVEL			
COURSE CONTENT (Objectives)	**TERMINOLOGY**	**UNDERSTANDING**	**CALCULATION**	**WEIGHT**
Comfort	3	5	1	18% (N=9)
Hygiene	6	7	2	30% (N=15)
Safety	3	5	3	22% (N=11)
Nutrition	3	3	9	30% (N=15)

Weight	30% (N=15)	40% (N=20)	30% (N=15)	50 Items

Several taxonomies have been developed to assist faculty in specifying the cognitive level of test items (see Table 2). Bloom's (1956) Taxonomy is familiar to many people. Another taxonomy which may relate more closely to course content in science was developed by Ebel (1965). A third taxonomy, developed by Williams and Haladyna (*Roid and Haladyna,* 1982), was specifically designed to assist in writing test questions. These taxonomies are guides that help instructors focus on the type of learning that is to be measured. In practice, the levels of the taxonomy may overlap; what may reflect comprehension for one student may be a knowledge level question for another. By targeting the test for the level of the course and the background of the students, however, the instructor can create an examination with whatever characteristics are appropriate for that course. It may be that in some beginning science courses, for example, it is necessary for students to learn a preponderance of facts, terminology, and classifications. Tests in these courses would reflect that emphasis, while in more advanced courses basic knowledge is assumed and the emphasis may be on analysis or prediction. The decision about which taxonomy or which portion of a taxonomy to use depends upon the content and the instructional objectives of the course.

Table 2

SPECIFYING COGNITIVE REQUIREMENTS

Bloom's Taxonomy	Ebel's Taxonomy	Williams & Haladyna
Knowledge Comprehension Application Analysis Synthesis Evaluation	Terminology Understanding Explanation Calculation Prediction Recommendation Evaluation	Reiteration Summarization Illustration Prediction Evaluation Application

■ *Increasing the Cognitive Level of Items*

A first step in writing items that measure more than recall of facts is to list content that students should know. This content could include concepts, events, facts or procedures, etc. Then ask, "Which discriminations should students be able to make, and what are some examples that could be used to illustrate the distinctions"? The

next step is to decide which cognitive level you want the item to measure. In the following section, two approaches to determining a cognitive level are presented using the Bloom and the Williams and Haladyna taxonomies. The reader can adapt this discussion to the use of the Ebel taxonomy listed in Table 2.

Using Bloom's Taxonomy:

A helpful chart based on Bloom's taxonomy lists key words associated with each level (see Figure 1). These key words can either be explicitly stated in the stem of the question or implied by the phrasing of the question.

Figure 1

Measuring Outcomes at Different Levels

If you want to measure:	Use these key words in the exercise:		
Knowledge	define	describe	identify
	label	list	match
	when	who	where
Comprehension	defend	discuss	distinguish
	explain	generalize	give example
	infer	how	why
Application	apply	demonstrate	discover
	modify	predict	relate
	show	solve	how
Analysis	compare	contrast	categorize
	break down	distinguish	identify
	illustrate	retake	select
Synthesis	combine	compile	explain
	integrate	organize	put together
	summarize		
Evaluation	appraise	argue	choose
	conclude	criticize	decide
	explain	interpret	support

-- adapted from Gronhund - 1970

Using the Williams and Haladyna Taxonomy:

This taxonomy was developed in response to criticisms that Bloom's taxonomy is primarily a classification system. The Williams and Haladyna system is designed to help item writers formulate questions rather than to classify them after they have been written. The premise of this system is that content can be classified in a three dimensional model according to (1) the content to be used (facts, concepts, or principles); (2) the use to be made of such content (the intellectual tasks of reiteration, summarization, illustration, prediction, application and evaluation); and (3) the response mode (selected--multiple choice or constructed--essay). The dimensions are described and summarized below.

• **Content Dimension:** Content is subdivided into facts, concepts and principles. ***Facts*** are defined as associations between names, other symbols, objects or locations; for example, "Springfield is the capital of Illinois" is a fact. ***Concepts*** are defined as classes of objects or events that are grouped together by virtue of sharing common defining attributes (i.e. thoughts or notions). An example of a concept would be, ***Mammals are a group of animals that share certain common characteristics such as giving milk to their offspring***. Items testing concepts might require students to (a) "identify examples from works by romantic poets, (b) name members of the brachiopod family, or (c) distinguish gases from liquids." ***Principles*** are defined as statements of relationship among objects or events (i.e. comprehensive assumptions) and can be communicated by if-then statements. An example would be, "If hot air rises, then cold air sinks." A subset of principles is ***procedure***, defined as the application of principles, or sequences of mental and physical activities used to solve problems, gather information, or achieve some defined goal. For example, given the principle "***i*** before ***e*** except after ***c*** or when sounded like ***a*** as in neighbor or weigh," then the procedure would involve asking the student to select or construct a word with ***ei*** or ***ie*** included in it.

• **The Intellectual Task (operational):** As devised by Williams and Haladyna, tasks correspond to cognitive levels (see Table 2). For example, the principle "A stitch in time saves nine" can be used according to the six different intellectual tasks outlined below.

Intellectual Task	Example
Reiteration	Recite verbatim "A stitch in time saves nine"
Summarization	Restate it in different words, such as "Make necessary repairs immediately to avoid having to spend a great deal more time making even more repairs later."
Illustration	Provide or identify examples of the rule in use, such as "Changing the oil in your car."
Prediction	Use the rule or principle to anticipate the consequences of certain acts, such as "Failure to change the oil in your car now will result in costly engine repairs later."
Evaluation	Employ the principle to make a judgment, such as "Is it better to change the oil now?"

Application	Application is the reverse of prediction. This would involve arranging conditions and initiating procedures to produce a desired outcome, such as "Given that you have to make costly engine repairs, what could you have done to avoid this"?

The last two intellectual tasks involve formulating possible courses of action rather than just selecting from options provided, so these two tasks mu st be used with the constructed rather than the selected, category of response mode in dimension three discussed below.

• **Response Mode:** The two response modes given are selected which involves choosing from among options (i.e. multiple choice questions), or constructed, which involves constructing and formulating short answers or essays.

The typology shown in Figure 2 shows how to take a fact, concept, or principle and make a multiple choice or essay question at the different intellectual levels from reiteration to evaluation. For example, the intellectual task of ***Prediction*** includes concepts, principles, and procedures that can be measured either in a selected (multiple choice) or a constructed (essay) mode. To measure the students' ability to make a prediction about a concept, the Taxonomy suggests that characteristics of an object be given and that students select other expected characteristics from a list (see Figure 2).

Figure 2

Williams and Haladyna Typology for Higher Level Test Items

Content/Task	Response Mode	
Reiteration:		
Facts, verbal description of concepts, principles, procedures, events, objects	Given a cue and a set of alternatives, select the exact replica of information previously learned.	Given a cue, construct an exact replica of information learned.
Summarization:		
Facts, verbal descriptions of concepts, principles, etc.	Given a paraphrased description, select the correct name from alternatives.	Given a paraphrased description, write the correct name for t he fact, concept, etc.
	Given the name, select the correct paraphrased description from alternatives.	Given the name, write a paraphrased description of the fact, concept, principle etc.
Illustration:		
Concepts, principles, procedures	Given the name, select a previously unused example of a concept, principle or procedure from alternatives.	Given the name, describe a previously unused example of a concept, principle, or procedure.
	Given a previously unused example of a concept, etc., select the correct name from a set of alternatives.	Given a previously unused example of a concept, principle, or procedure, name it.
Prediction		
Concepts	Given characteristics of an object, select other expected characteristics from a list.	Given some characteristics of an object, list other expected characteristics.
Principles	Given a previously unused description with	Given a previously unused description with

	antecedent conditions of a relationship embedded, select the most likely consequences.	the antecedent conditions of a relationship embedded, describe the most likely consequences.
	Given a previously unused description, of an outcome, select the antecedent conditions to achieve the outcomes.	
Procedures	Given a previously unused description of a situation in which a procedure is used, select the likely outcome.	Given a previously unused description of a situation in which a procedure is used, describe the most likely outcome.
	OR, give the outcome and select the procedure used.	OR, give the outcome and describe the procedure used.
Application		
Concepts, principles, procedures		Given a previously unused case and a desired outcome, arrange or perform the antecedent conditions to achieve that outcome.
Evaluation		
Principles, procedures		Given a description that calls for a judgment, name the most desirable action and describe the criteria on which your decision was based.

■ *Planning the Items*

When the decision has been made about what content to test at which cognitive levels, the test developer must select the appropriate types of questions to use and must also determine the range of difficulty for the test.

Selecting Item Format:

The choice of item format (essay, short answer, multiple choice) is related to the nature of the content, the test objectives and the time available to write items. Thorndike and Hagen (1969) have summarized the advantages of various essay and objective test formats. Positive attributes have been designated by plus signs in the table below, while distinct disadvantages have minus signs.

Table 3

RELATION OF ITEM TYPE TO TEST OBJECTIVES

FACTOR	ESSAY or ORAL	SHORT ANSWER	MULTIPLE CHOICE
Ability to organize	++	+	–
Discourages bluffing	—	–	++
Potential diagnostic value	—	+	++
Easily and reliably scored	—	+	++

Takes little time to prepare	+	+	–
Measures higher mental processes	++	–	++
Broad content sampling	—	+	++
Measures application	++	+	++
Adequate sampling of objectives	—	+	++

Multiple choice tests are useless as measures of the ability of students to organize ideas, but they give reliable estimates of the breadth of student learning. In survey courses, large areas of content are covered and multiple choice tests are better suited to measure that range. Essay tests restrict the amount of material that can be assessed, and essay questions are often too broadly construed to measure fine distinctions in understanding. The choice of question format depends upon the purpose for testing and the time available for test construction and scoring. High quality multiple choice items are difficult to construct but easily and reliably scored while the converse is true for essay tests. Students need to write to organize and clarify their thinking; they can have their knowledge and critical thinking ability tested as well by using a combination of essay, multiple choice and performance measures.

Determining Item Difficulty:

Tests are criterion-based when the purpose is to determine how many students have mastered certain content, or they may be norm based when the purpose is to differentiate among students. Mastery tests require several questions about each content subarea in order to make a judgment about mastery. The difficulty of the questions is determined by the content and the judgment of the instructor about the required standard for mastery. Norm based tests consist of a collection of questions that represent the content, but they are constructed to spread scores out on an ability continuum rather than to group scores that represent categories of knowledge or skills. Many university course examinations are norm referenced, and the following discussion is related to this type of examination.

If the goal of a test is to differentiate among students, then items should be constructed with a medium level of difficulty and a narrow range of difficulty. Lord (1952) published the following levels of difficulty to maximize test reliability.

Table 4

GUIDE FOR PREPARING TESTS WITH DIFFERENT ITEM FORMATS	
ITEM FORMAT	IDEAL AVERAGE DIFFICULTY

Completion and short answer	50*
Five-response multiple choice	70
Four-response multiple choice	74
Three-response multiple choice	77
True-false	85*

* Expect a student of average ability to answer about one-half the items correctly. (Lord, 1952). Difficulty is defined as the percent correct.

The appropriate range of difficulty is related to the intercorrelation of the items and the purpose of the test. Tests with low item correlations (tests in which different groups of students tend to get different items correct) will have most items centered at a difficulty level of .50 for maximum discrimination among all examinees. For tests with high item correlations, such as vocabulary tests, the item difficulties are spread out over a range in order to maximize discrimination. If the purpose of the test is to classify the examinees into two groups, the bulk of the items should be centered at the passing standard. For classroom testing, where discrimination is important at several levels, the item difficulties should be spread in a rectangular pattern around the ideal average difficulty. Thus, for a five response multiple choice test, the average difficulty would be 70 percent, and an equal number of questions would be placed at difficulties ranging from about 50 percent to 90 percent correct.

In order to know how difficult an item is, an experienced instructor might use professional judgment initially and then use an item analysis in which the difficulty and discrimination of an item is calculated. Instructors often create item banks in which the item statistics are cumulated so that an accurate estimate of difficulty can be made. These banks may be kept on card files or on personal computers. It is a good practice to identify each item with a unique number that represents the content and to save each item and its statistics as a separate file using the item number as a file name. The item statistics can be deleted from the test file once the test is developed. Some commercial test construction software is available, but wordprocessing software such as Wordperfect is easily adapted for this purpose.

Plan the difficulty of the items in conjunction with the objectives for the course. The harder items have more influence on the total score. If these items are concentrated in one content area, that area will unequally weight the total score. Avoid tests that are too difficult for even the best students; these tests destroy motivation. Good questions to ask are:

- Should difficult items be from all content areas?
- If one area is more difficult, should it include fewer items to balance its effect on total test score?

Essay questions can be made more difficult if students are required to analyze a concept or an event rather than to describe one. To increase the difficulty level of the multiple choice items:

- Use compound response (***a*** and ***c*** responses are correct but not ***b***)
- Increase the cognitive level
- Make options more homogeneous

Avoid trick questions or obscure facts; they reduce test reliability.

Planning the appropriate difficulty increases the reliability of the test and reduces the effect of guessing. On a well constructed test, a student should be expected to receive the same score if the test were given on two different occasions without additional instruction: such a test is considered to be a reliable measure. As guessing levels increase, reliability decreases.

Reducing Guessing:

The difficulty level of an item is the percent of the examinees who correctly answer an item. The difficulty of multiple choice items is related to guessing in the following way. If there are four alternate responses to an item, and the difficulty level is .70, then the 30 percent who failed the item would be distributed equally over the three incorrect responses (assuming each response is equally plausible). Therefore, the possibility of guessing any single response including the correct one, is 10 percent. Thus, while 70 percent of the students correctly responded, it is likely that 60 percent actually knew the correct answer. The more difficult the item, the greater the effect of guessing. This concept is illustrated in Table 4.

Table 4

CONSIDER THE GUESSING RATE		
70% Pass	**30% FAIL** **foil 1 - 10%** **foil 2 - 10%** **foil 3 - 10%** **Foil 4 (correct response)**	**60% knew the answer**
30% Pass	**70% FAIL** **foil 1 - 23%** **foil 2 - 23%** **foil 3 - 23%** **foil 4 (correct response)**	**7% knew the answer**

Avoiding Gender and Ethnic Bias:

In some course content, bias is inherent and students are taught to recognize and evaluate it. Examinations that include bias that is part of the course content are certainly appropriate. The goal in test construction is to reduce measurement error that results from interpretations of questions that are related more to sex, racial or ethnic

background than to the skill to be assessed. Some aids in reducing bias in test items are listed below:

- Substitute plural pronouns to eliminate he/she usage.
- Avoid sexual or ethnic stereotypes in occupations, roles, etc.
- Use situations, places, and experiences that are common to the whole population.

II. Writing and Evaluating Multiple Choice and Related Items

■ *Using Guidelines to Construct Multiple Choice Tests*

A content valid test measures what it is intended to measure — the content as represented by the test blueprint. A reliable test yields the same scores over repeated administrations. A test that is not reliable cannot be valid. To ensure that tests are both valid and reliable, adequate planning and careful item construction are necessary. In this section the construction of test questions is discussed. Guidelines for constructing test items are reviewed, and different item formats are presented.

The problems with most test items relate to reading load, clues embedded in the questions or ambiguity in the phrasing of test items. The use of unnecessarily difficult vocabulary can confuse students who otherwise could respond adequately. Tests with poorly constructed items are not accurate measures of what students know and may not rank students in the same way as a well constructed examination. The following guidelines provide examples of items with common flaws.

I. Reducing Errors Due to Reading Difficulty or Reading Speed

•Avoid writing items with an unnecessarily difficult vocabulary.	
Poor 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.	***Better*** Frequent use of sprays, oils, and antiseptics in the nose during acute colds may result in A. spreading the infection to the sinuses. B. damage to the olfactory nerve. C. destruction of white blood cells.

•Avoid repeating words.	
Poor Entomology is A. the study of birds. B. the study of fish. C. the study of insects.	***Better*** Entomology is the study of A. birds. B. fish. C. insects.

•Eliminate unnecessary words.	
Poor There were many different theories about the origin of mankind. The man associated with the theory of evolution was A. Darwin. B. Galileo. C. Freud.	***Better*** The man associated with the theory of evolution was A. Darwin. B. Galileo. C. Freud.

•State the problem in the stem.	
Poor Abraham Lincoln A. chopped down the cherry tree. B. signed the Declaration of Independence. C. wrote the Emancipation Proclamation.	***Better*** The president who wrote the Emancipation Proclamation was A. George Washington. B. Thomas Jefferson. C. Abraham Lincoln.

II. Avoiding Giving Clues in the Question

•Watch for grammatical clues.	
Poor A long, winding, gravel crest of glacial origin is a A. kame. B. ridge. C. esker.	***Better*** A long, winding, gravel crest of glacial origin is a/an A. kame. B. ridge. C. esker.

•Avoid using clues in the stem that give away the correct response.	
Poor When linking two clauses, one main and one subordinate, one should use a A. coordinate conjunction such as *and* or *so* B. subordinate conjunction such as *because* or *although*. C. preposition such as *to* or *from*. D. semicolon.	***Better*** When linking two clauses, one main and one dependent, one should use a A. coordinate conjunction such as *and* or *so*. B. subordinate conjunction such as *because* or *although.* C. preposition such as *to* or *from*. D. semicolon.

•Avoid making the longest response correct.	
Poor The term "side effect" of a drug refers to A. additional benefits from the drug. B. the chain effect of drug action. C. the influence of drugs on crime. D. any action of a drug in the body other than the one the doctor wanted the drug to have.	***Better*** The term "side effect" of a drug refers to A. additional benefits from the drug. B. the chain effect of drug action. C. the influence of drugs on crime. D. any unwanted action of the drug.

•Use plausible options.	
Poor Who discovered the North Pole? A. Christopher Columbus B. Ferdinand Magellan C. Robert Peary D. Marco Polo	***Better*** Who discovered the North Pole? A. Roald Amundsen B. Richard Byrd C. Robert Peary D. Robert Scott

• Use **none of the above** sparingly. It is best to use this only when the keyed answer can be classified unequivocally as right or wrong, for example on tests of spelling, mathematics, and study skills. Be certain that this is not always the correct answer.

***•Avoid the use of* all of the above.**	
Poor Which of the following factors must be considered in computing basal energy requirements? A. Age B. Height C. Weight D. All of the above	***Better*** Which of the following factors must be considered in computing basal energy requirements? A. Weight only B. Age only C. Height and weight only D. Age and weight only E. Age, height, and weight

• Vary the position of the correct response.

III. Avoiding Ambiguity in the Question

•Include only one correct or best answer.	
Poor A color-blind boy inherits the trait from a A. male parent. B. female parent. C. maternal grandparent. D. paternal grandparent. E. remote ancestor.	***Better*** A color-blind boy most probably inherited the trait from his A. father. B. mother. C. paternal grandfather. D. paternal grandmother.

•***Write options that have parallel grammatical construction.***	
Poor An electric transformer can be used A. for storing up electricity. B. to increase the voltage of alternating current. C. it converts electrical energy into mechanical energy. D. alternating current is changed to direct current.	***Better*** An electric transformer can be used to A. store up electricity. B. increase the voltage of alternating current. C. convert electrical energy into mechanical energy. D. change alternating current to direct.

•***Place blanks near the end of the sentence.***	
Poor The _____ is a long haired wild ox found in Tibet.	***Better*** The long haired wild ox found in Tibet is called a _____.

•***Avoid negatives in the stem.***	
Poor Which one of the following is not a safe driving practice on icy roads. A. Accelerating slowly B. Jamming on the brakes C. Holding the wheel firmly D. Slowing down gradually	***Better*** All of the following are safe driving practices on icy roads EXCEPT A. accelerating slowly. B. jamming on the brakes. C. holding the wheel firmly. D. slowing down gradually.

A tip you might consider when you write multiple choice questions is to try out question ideas in a short answer format first. Then use common errors that students make for the foils in the multiple choice question.

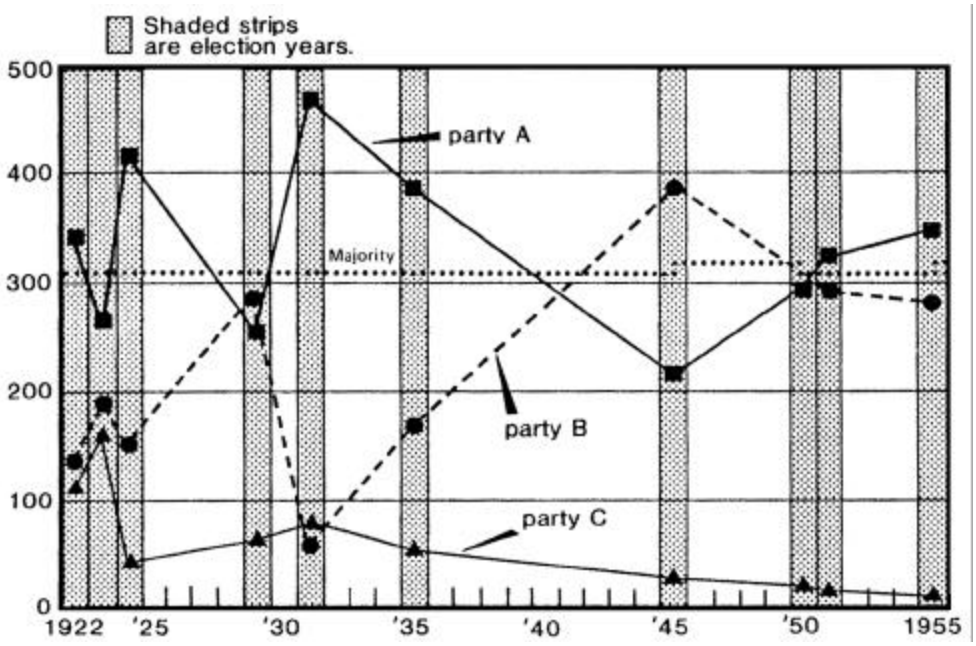
■ *Selecting Question Formats*

In this section of the handbook, there are examples of different types of question formats that can be used for machine scoreable examinations. Several varieties of multiple choice questions are illustrated as are alternate response, keylist, and incomplete response questions. Be certain to give clear directions so the student will understand the task.

Multiple Choice:

• *Give an example of a phenomena and ask students to identify it by analyzing the information:*

The graph above represents the political composition from 1922 to 1955 of which of the following?



A. German Bundestag
B. French National Assembly
C. Italian Chamber of Deputies
D. British House of Commons

Educational Testing Service, 1963

This question illustrates a political system with two major and one minor political party. Students who can interpret the graph will recognize that the majority party changed in the 1945 election and that fact combined with knowledge of the four party systems should lead students to (D) the correct response.

•Require students to compare alternatives

Which of the following economic policies is most likely to produce effects that are inconsistent with the effects of the others?

A. Liberalizing installment buying plans
B. Increasing public works spending
C. Lowering interest rates
D. Increasing expenditures for industrial plant and equipment
E. Increasing taxes

Educational Testing Service, 1965.

•Combine two elements in the stem and in the responses.

How has the rate of unemployment in the United States been affected by the size of the American foreign trade deficit since 1980 and why?

A. It has decreased because American business now recognizes the economic value of hiring American labor.
B. It has decreased because tariffs on foreign goods have raised the price of these products to an extent where American goods are less expensive.
C. It has increased because cheap foreign labor makes foreign goods less expensive.
D. It has remained the same because the trade deficit has inspired a "buy American-made products" that has offset the initial decline in employment.

Editorial Manual for Item and Test Development, Instructional Research and Practice, University of South Florida, 1986.

•Use scenarios or problem situations

A bag containing a solution "A" was dropped into a beaker of water. After five minutes it was found that water was moving into the bag containing solution "A".

1. The movement of water into the bag is an example of the process called

   A. diffusion.
   B. osmosis.
   C. active support.

2. Water will continue to move into the bag until the number of molecules of

   A. solution "A" is equal on both sides of the bag.
   B. solution "A" is greater outside the bag than inside the bag.
   C. water is greater inside the bag than outside the bag.
   D. water is equal on both sides of the bag.

3. What will occur when this experiment reaches equilibrium?

   A. All movement will cease.
   B. The concentration of solution "A" will be the same on both sides of the bag.
   C. For every molecule of solution "A" that leaves the bag, another molecule will enter the bag.
   D. For every molecule of water that leaves the bag, another molecule of water will enter the bag.

Adapted from: Carlson S., *Creative Classroom Testing*. Educational Testing Service, Princeton, N.J. 1985.

•Try some ill-structured reasoning questions.

Reasoning questions can be tightly structured deductive problems as are syllogisms in logic (If p then q; not p then not q). Ill-structured problems are similar to scientific thinking problems and to logical reasoning questions; they differ in that they require the examinee to structure the assumptions needed to solve the problem (Ward, 1983).

Well structured questions require students to understand and manipulate the information usually with the aid of an algorithm; there is one correct answer, and the student must apply the correct procedure to find the answer. Ill-structured questions do not provide all the information necessary to solve a problem and are similar to some questions on examinations such as the Law School Admissions Test or the Medical College Admissions Test. Students may be given the facts of a case, the decision to be made, and the rules that apply. Then students must respond to a series of questions that would arise in the decision making process. If all the facts are present and the rule is unambiguous, then the question is tightly structured. If not all the facts are available or the rules may be interpreted in more than one way, then the problem is ill-structured; often more than one response may be correct. If these get too complicated, an essay format is best.

To write these questions, the instructor identifies a problem and asks what assumptions are required to link an argument to a conclusion or what inferences might plausibly be drawn from a finding. Sample questions that could be posed are listed below and would follow a case study:

1. Which of the following, if true, is most likely to explain the finding above?

2. One possible explanation of the finding is A. To evaluate this explanation, it would be useful to know each of the following EXCEPT:

3. One possible interpretation of the finding is X. One could argue against this interpretation most effectively by pointing out:

•Test student ability to reason logically in the subject area.

Questions can be written in most subject areas that measure students' ability to apply their knowledge of logical reasoning. These logical fallacies include such things as circular logic, red herrings, faulty cause and effect, non sequiturs, begging the question, etc. (See Wyrick, 1990 for more examples.) Instructors may have to teach students some of these reasoning techniques, but the reasoning can be applied to specific content areas and its use avoids the pitfalls associated with many higher order multiple choice questions in which students tend to argue that any answer is potentially correct.

1. The format of these questions varies. One method to write questions is to write statements that contain reasoning errors and ask students to identify the type of error as in:

   Both the French and the Russian revolutions were followed by periods of literary rebellion. Clearly, political revolt results in revolt in the arts.

   What leads you to believe that the argument in this statement is illogical?

   A. The facts given in the sentences above have been proven by history.
   B. It is a faulty cause and effect relationship between political revolt and literary rebellion.
   C. This proves that a revolution will eventually bring about a literary rebellion.
   D. Literary figures were eager to demonstrate their support of the revolution.

2. Another format requires students to identify either the statement that contains a fallacy or one that does not such as:

All of the following contain fallacies in reasoning except:

A. You should use Zippy Oil because 9 out of 10 mechanics prefer it.
B. Professor Jones always wears old sneakers and a vest. He must not know very much about history if he cannot learn to dress.
C. The college catalog states that an "A" has a value of four quality points.
D. History proves that wealthy nations control the food supply.

Questions adapted from the *Item Specifications for the College Level Academic Skills Test*, Florida Department of Education, 1990.

•Combine responses to increase the need for discrimination.

Which of these yield a basic solution in water?

1. NaOH 2. NH3 3. HCl

A. 1. only
B. 2. only
C. 3. only
D. 1. and 2.
E. 2. and 3.

Alternate Response

A common version of the alternate choice is the true-false item. Many item writers argue that true-false questions are poor because it is difficult to write statements that are clearly true or false in all conditions. Even if qualifiers such as "always" or "never" are included in the statements, the better students often find plausible reasons why true-false items have no clearly correct answer particularly for test items that measure higher order thinking skills. For this reason, it is particularly important to use attribution (According to X theory, what is ...) in constructing questions or to use other qualifiers that constrain the student response. An example of a well constructed true-false item (Wesman, 1976) that can measure reasoning is:

"If a square and an equilateral triangle are inscribed in the same circle, the side of the square is longer than the side of the triangle." F

A variant of the true-false question that is considered useful for measuring higher order thinking skills is the cluster or multiple true-false type. Multiple true-false questions include several separate questions about a single topic or concept. Research on this format (Frisbie, 1990) indicates
that multiple true-false tests may be more reliable than standard multiple choice tests if certain guidelines are followed. These guidelines are:

• Use at least a ratio of 2/3 multiple choice to true-false items for the same time period.
• Within each cluster, use at least 50% false answers.

• Score each question separately; though you may want to weight a multiple true-false section of a test less in order to balance the increased number of questions from this format with the fewer points possible from a standard multiple choice section of a test.

An example of a cluster type multiple true-false item (Frisbie, 1990) is shown below:

The set of scores 4, 5, 5, 7, 8, 9, 9, 15 has

(T) 1. a median of 7.5.

(F) 2. a symmetric shape.

(T) 3. more than one mode.

(T) 4. a larger mean than median.

(F) 5. a mean of 7.0.

Some highly structured alternate response items can assess students' understanding of functional relationships or principles. Ebel (1982) illustrated this form:

The density of ice is (1) greater or (2) less than that of water?

These items minimize reading load, allow broad content coverage and can be more precisely stated. Wesman (1976) adds this caveat: Students have a 50% chance of correctly responding by guessing. It therefore takes more items to achieve the reliability of multiple choice examinations.

Key Lists

Key lists allow students to differentiate among aspects of related subject matter. The question below requires students to apply their knowledge about political systems in new situations that were not specifically taught. Keylists can take many forms from simple classifications to comparisons of functions in which the keylist may include combinations of processes that might be related to particular outcomes.

Examples of key lists are given below:

- For each of the statements about characteristics of political systems in questions 1-6, use the letters below if the statement describes a:

  A. high-consensus, liberal constitutional system.

  B. consociational system.

  C. "stalled" system.

  D. "bureaucratic" system.

  E. mobilization system.

  1. The political system most likely to have foreign policy goals directed toward changing the structure and norms of the international system.
  2. The political system most likely to provide opportunities for foreign intervention.
  3. The political system where the rulers are likely to be the major beneficiaries of defense expenditures.
  4. The political system most likely to increase national capability rapidly.

5. The political system least likely to be actively involved in international politics.
6. The political system in which the potential for expansion of governmental capability in a short time is the greatest.

- Below are given a series of hypotheses, each of which is followed by numbered items which represent data. After each item number on the answer sheet blacken space if the item
  - A. directly helps to prove the hypothesis true.
  - B. indirectly helps to prove the hypothesis true.
  - C. directly helps to prove the hypothesis false.
  - D. indirectly helps to prove the hypothesis false.
  - E. neither directly or indirectly helps to prove the hypothesis true or false.

  1. In a partial vacuum....
  2. Scientists now believe...

Incomplete Response

This form is most effectively used when it is important for students to recall rather than recognize the correct answer, or as in the case of some mathematics problems, when the instructor wants the student to calculate an answer without the assistance of possible correct answers. For example:

When the square root of 21 is computed, what number should appear in the second decimal place?

A. 2

B. 4

C. 5

D. 8

Short Answer

These questions are written as completion items or direct questions or statements to which students are directed to respond. Completion items (fill-in the blank) tend to be the most difficult to write clearly. If they are used, the blanks should be near the end of the sentence and the question needs to be precisely stated:

Poor: How many _____ are in the universe?

Better: The number of planets in our solar system is _____?

Short answer items require students to recall information, but they also tend to elicit rote answers and/or answers that are difficult to score. Instructors are faced with the problem of allocating partial credit because answers are too brief, vague, or otherwise imprecisely stated. An illustration of the ability of multiple choice items to more precisely measure understanding is shown in a test for 325 Iowa high school physics students (Lindquist, 1936).

1. What is the heat of fusion of ice in calories? (*80 calories per gram*)

(Answered correctly by 75 percent of the students.)

2. How much heat is needed to melt one gram of ice at 0 degrees C.? (*80 calories*)
(Answered correctly by 70 percent of the pupils.)

3. Write a definition of heat of fusion. (*Amount of heat required to change a given quantity of ice at 0 degrees C. to water at O degrees C.*)
(Answered correctly by 50 percent of the students.)

4. In which of the following situations has the number of calories exactly equal to the heat of fusion of the substance in question been applied?

A. Ice at 0 C. is changed to water at 10 C.
B. Water at 100 C. is changed to steam at 100 C.
C. Steam at 100 C. is changed to water at 100 C.
D. Frozen alcohol at -130 C. is changed to liquid alcohol at -130 C.
(Answered correctly by 35 percent of the students.)

■ *Analyzing the Test*

Item analysis helps to determine the adequacy of the items within a test as well as the adequacy of the test itself. The results of an item analysis provide information about the difficulty of the items and the ability of the items to discriminate between better and poorer students.

• **The index of difficulty** is defined as the proportion of students who correctly answer an item. The average difficulty of a test is the average of the individual item difficulties. For maximum discrimination among students, an average difficulty of .60 is ideal.

Example: If 243 students answered item no. 1 correctly and 9 students answered incorrectly, the difficulty level of the item would be 243/252 or .96.

• **The index of discrimination** is a numerical indicator of how the poorer students answered the item as compared to how the better students answered the item. The scores are divided into three groups with the top 27% of the scores in the upper group and the bottom 27% in the lower group. The number of correct responses for an item by the lower group is subtracted from the number of correct responses for the item in the upper group. The difference in the number correct is divided by the number of students in either group. The process is repeated for each item.

Example: Sixty students take a test. The top 16 scores and the bottom 16 scores are the upper and lower groups. For item no. 1, twelve of the sixteen students in the upper group answered the item correctly while seven students in the lower group answered correctly. The index of discrimination for

$$\frac{12 - 7}{16} = .31$$

item no. 1 would be calculated as follows:

For a small group of students, an index of discrimination for an item that exceeds .20 is considered satisfactory. For

larger groups, the index should be higher because more difference between groups would be expected. The guidelines for an acceptable level of discrimination depend upon item difficulty. For very easy or very difficult items, low discrimination levels would be expected; most students, regardless of ability, would get the item correct or incorrect as the case may be. For items with a difficulty level of about 70 percent, the discrimination should be at least .30.

• **The point biserial correlation** is a coefficient that represents the 1, 0 (correct, incorrect) correlation between the item response for each student and the student's total test score. Conceptually, the point biserial is similar to the discrimination index, but the point biserial includes the data for the middle group of students from the score distribution. One would expect that higher scoring students would get each item correct. If lower scoring students got a particular item correct more often than did higher scoring students, then that item would have a negative point biserial correlation. As a guideline, items with point biserials below .30 should be reviewed for possible deletion if the items are of moderate difficulty.

■ *Reliability*

Reliability may be defined as the degree of consistency or stability between two measures of the same thing. When two measures are not available, reliability is estimated by the degree to which the items correlate with the total test score. When the items are of equal difficulty, the KR_{21} formula shown below may be used. When items have different levels of difficulty, the KR_{21} will underestimate the reliability, and OIR uses the computationally more

$$r_{xx} = \frac{N}{N-1}\left(1 - \frac{x(N-x)}{NS_x^2}\right)$$

complex KR_{20} formula to estimate reliability.

where: r_{xx} = the KR_{21} estimate of reliability

N = the number of items in the test

x = the mean score of the total test

S_x^2 = the variance of the total test

Reliability coefficients are related to the level of discrimination of the test items and to the number of test questions. Longer tests with high item discriminations are more reliable measures. For most instructor-made tests with about 50 items, a reliability coefficient of .75 is respectable. Standardized examinations would be expected to have reliabilities in the high .80's or .90's.

After the test is constructed and administered, the adequacy of the test as a measure of student achievement can be evaluated. The individual item contribution to test reliability can be assessed by comparing the item difficulty, discrimination and point biserial correlation. For a well constructed item, the student responses should be evenly distributed across the item foils (incorrect responses) with the correct answer receiving the highest

percentage of responses. A concentration of incorrect responses in a single foil may indicate problems in the wording of the items. Effective items also discriminate between high and low scoring students. In the sample analysis below, the first item is too easy to discriminate. The second item operates very well, and the third item should be revised to improve the a and b foils as well as its discrimination. In this item, it is likely that good students chose foil **e** because nearly equal numbers of high and low scoring students correctly responded (see Table 6).

Table 6

ANALYZE THE TEST RESULTS

	ITEM RESPONSES					DIFFICULTY	DISCRIMINATION	POINT BISERIAL
	a	b	c	d	e			
1.	2	4	243+	2	1	0.96	0.00	0.00
2.	26	21	28	149+	23	0.59	0.51	0.59
3.	0	0	5	219+	28	0.86	0.09	0.10

If an item is miskeyed or misleading, the instructor may wish to give students credit for the item or to delete the item from scoring. There is generally less confusion if all students are given credit than if the item is initially scored as correct and then rescored as incorrect. An exception to this procedure might be if the test has criterion referenced sections in which a specified number of items correct indicate student mastery of the content. In that event, it may be better to delete the item from scoring.

III. Writing and Scoring Essay Questions

Essay examinations are useful tools to evaluate students' thought processes if they are constructed to elicit more than a rehash of material presented in class. Unfortunately, essay examinations are often subjectively and inconsistently scored. These problems may be related to the construction of the essay items, reader bias, or to the scoring procedure. In this section, guidelines for improving the reliability and validity of essay scores are presented and examples of essay topics are critiqued. The following suggestions help to improve essay examinations:

1. Frame questions so that the task is clearly defined.
2. Specify the value and approximate time limit for each question.
3. Do not employ optional questions, but do use several essay questions or a combination of essay, short answer, and multiple choice questions.
4. Write a trial response to each essay question.
5. Prepare a tentative scoring key in advance.
6. Alert students if penalties for spelling or grammar are given.
7. Evaluate essays anonymously by mixing papers and masking names.
8. Scan longer essays to presort into stacks of high, medium and low quality and then reread to assign marks.
9. Score all answers to one question before scoring the next question.
10. Decide whether essays are to be scored holistically or analytically. Holistic scores are awarded as a global measure of overall quality, i.e. a single mark is given on a 1 - 5 (A - E) or perhaps a 10 or 12 point scale. An analytic score is the sum of several different ratings of features about an essay. An example of analytic scoring is found on page 29.

When a combination of essay and multiple choice items are included in a test, the calculation of the exam score can be handled in several ways. Perhaps the easiest method is to assign weights to each section of the test. If the essay portion is worth 50 percent and the multiple choice worth 50 percent, the essay might be rated on a scale from 0 to 10 points, either holistically or analytically. The total number of essay points could be multiplied by 5 making a range of 0 to 50 possible points for the essay. The difficulty in using a single scale from 0 to 50 or more points to initially rate essays is that it is extremely difficult for a reader to reliably differentiate papers that receive, for example, 36 or 38 points. A second reading might well reverse the two scores.

The following discussion about essay examinations is particularly helpful in evaluating essay questions. It is adapted from a bulletin published at Michigan State University.

The Writing of Essay Questions

It is commonly recognized that the reading and marking of essay examinations is a difficult and time-consuming process; it is not always recognized that the preparation of essay questions requires time and effort. The weaknesses which have frequently characterized the essay question have, in part, resulted from a lack of awareness of or concern about these difficulties. Overemphasis on factual information and on memorized extracts of textbooks and lectures, ambiguity in meaning and an associated ambiguity of response, and variations in the standards used to judge student performance result in difficulties in reading and in unreliable marking. Essay questions can be markedly improved by conscious attention to the identification and elimination of these and other deficiencies.

Perhaps the most general suggestion for improving essay testing is that the task set for the student should first be clearly understood by the teacher and then clearly presented to the student. It is unreasonable to present the student with the combined tasks of reading the instructor's intent in the question and meeting his expectation as to the answer. On the other hand, to make an essay question highly specific and detailed is to reduce it to a type of objective test item involving a filling in of blanks implied in the question. The poser of essay questions must tread the path between highly general questions to which no adequate answer can be given in the time permitted — even assuming the meaning is clear — and highly specific questions which involve word or phrases answers. The question must state precisely what is wanted, and it should contain enough clues that students can readily infer the direction their answer should take. It is also desirable that the question should state or imply the approximate length and complexity of the desired response. Some examples may serve to clarify this and other points.

Example 1: Account for the simultaneous rise of Protestantism and Capitalism in the 16th and 17th centuries.

chosen. One student described the large impersonal forces of nationalism-imperialism-communism; another considered purely political events which took the form of diplomatic documents; and still another listed those happenings of the few weeks preceding the outbreak of hostilities. Each treatment is an

This question was probably intended to stimulate students to organize their knowledge of these developments and to seek for any relationship which seemed to obtain. If so, the question demonstrates a serious weakness. Some students might undertake to show that there was a significant relationship and that there were elements in Protestantism which caused, or at least conditioned, the form modern capitalism took. Other students might premise that the two movements had only a temporal relationship. Thus, at least two and possibly more theses are possible, each of which would require an essay of some length for adequate development. For this interpretation the question is too general and the answers will be difficult to read and to mark. Perhaps, however, the intent was only that the students reproduce verbatim or paraphrase a discussion of this point as presented in the text or in a lecture. In the latter interpretation a well-defined standard exists for marking, but the task presented is of less significance. A more satisfactory question involving something of the same considerations would seem to be found in:

Indicate the principal evidence upon which Max Weber based his theory of the Protestant Ethic and how the conclusions reached by Weber differ from other major interpretations of the matter.

If the student's own point of view is desired, further specification is required. Since the student's point of view may be based more on his beliefs than on the evidence, there arises some difficulty in fairly judging the personal response.

Example 2: Describe the origins of World War I.

The question was intended to test the knowledge of the complex of events in Europe from 1870 to 1914 and their role as causal factors in promoting war. It failed as a measuring device because of the ambiguity of the wording. Neither the word "describe" nor the word "origins" is well

appropriate response, but the range of interpretation complicates the problem of judging between the levels of performance. The word *describe* does not indicate precisely what the student is expected to do. Should he present an objective catalog of facts? Should he attempt to interpret their meanings? Should he attempt any sort of value judgement? The following question eliminates some of these ambiguities:

What were the principal diplomatic events in Europe between 1890 and 1913 which contributed directly to the outbreak of World War I?

This question unequivocally calls for selection and presentation of certain items of knowledge. The extent to which it requires thought and judgement depends upon what has gone on in the classroom. It is likely that this particular question will not evoke everything that the instructor had in mind. If so, other questions should be posed.

Example 3: Comment on the significance of Darwin's *Origin of Species*.

Again here is a question which allows such latitude as to give students an impossible problem of selection and teachers an equally impossible task of appraisal. The intent was apparently to give students sufficient range in which to display their mastery of the material, although it may only have involved recall of points made by the instructor or textbook author. However, one student might analyze Darwin's conclusions in the light of present knowledge, another might place the book in the context of 19th-century intellectual history, while a third might show its impact on 20th-century thinking outside the scientific fields. While freedom and individuality in response are desirable, the measure afforded here overflows a one- or two-hour examination period. These time limits require that students be presented with a clearly problematic situation, the solution to which calls for one line of reasoning. For example, the same general value of the question might be achieved by revising it to say:

Darwin in his *Origin of Species* emphasized that natural selection resulted in the survival of the fittest. To what extent has this been supported or refuted by subsequent biological research.

Example 4: Germany, France, England, and Russia were all partly to blame for the outbreak of war in 1914. Tell the substance of the final diplomatic message sent by each of the four countries to the other three before hostilities commenced. Who was the official who signed each message?

Such a question is made to order for the student with a memory. The utility of such a highly specific question is debatable. To answer it, students would normally be required to spend more time than the results are worth, either as a sample of an individual's knowledge or as a test of his abilities to organize and synthesize.

Illustrations of Questions Requiring Various Levels of Response

In framing an essay question one should decide on the ability or the level of response that is desired and so phrase the question as to make this clear. The following questions have been selected because they are reasonably clear in this respect, although they are not necessarily model questions otherwise.

I. Recall

A. Simple recall
–What is the chemical formula for hydrogen peroxide?

B. Selective recall in which a basis for evaluation or judgement is suggested
–Who were the three most significant men in the 19th century with respect to their ability to describe nature.

II. Understanding

A. Comparison of two things on a single designated basis
–Compare Renaissance writers with writers of the 19th century with respect to their ability to describe nature.

B. Comparison of two things
–Compare the French with the Russian Revolution.

C. Explanation of the use or exact meaning of some phrase or statement
–The Book of John begins with the statement, "In the beginning was the Word...." From what philosophical system does this derive?

France is still torn along lines established during the French Revolution?

D. Summary of some unit of the text or of some article
–Why does the author say he believes modern

E. Statement of Aim — author's purpose in his

selection or organization of material
–In his Symphony No. 6, what was Beethoven's purpose in varying the orthodox form of a symphony?

III. Application: It should be clearly understood that whether or not a question does elicit application depends on the educational experience which precedes it. If an analysis has been taught explicitly, then a question involving that analysis becomes simply a matter of recall.

A. Causes or effects
–Why may too frequent reliance on penicillin for treatment of minor ailments eventually result in diminished effectiveness of penicillin treatment against major invasion of body tissues by infectious bacteria?

B. Analysis (It is best not to use the word in the question)
–Why was Hamlet torn by conflicting desires?

C. Statement of relationship
–Penalties for sexual crimes in America are typically more severe than for similar crimes of a non-sexual nature. Why is this so?

D. Illustrations or examples of principles
–An individual robbed a bank in 1932 and then fled the state. He returned to the state only once between 1932 and 1942, and that was in 1938. In 1942 he was arrested in a neighboring state. On what grounds should he or should he not have been extradited?

E. Application of rules or principles in specified situations
–Would you weigh more or less on the moon? On the sun? Why?

F. Reorganization of facts
–Pavlov found that sometimes dogs he had conditioned to salivate when a bell rang would fail to do so. How do you account for this?

IV. Judgement

A. Decision for or against
–Should members of the Communist Party be allowed to teach in American colleges? Why?

B. Discussion
–Discuss the likelihood that four-year privately supported liberal arts colleges will gradually be replaced by junior colleges and state universities.

C. Criticism as to the adequacy, correctness, or relevance of a statement
–The discovery of penicillin has often been called an accident. Comment on the adequacy of this explanation.

D. Formulation of new questions
–What should one find out in order to explain why some students of high intelligence fail in school while others succeed?

In preparing any question for an essay examination, the teacher ought to try to anticipate what responses it will probably elicit. Not only will such a procedure point out weaknesses in the question, but it will suggest how answers should be scored. Writing out an answer is a recommended practice because it forces the questioner to make explicit what he hopes to elicit. He can then better judge the adequacy of the questions. Consider the following question:

What are the principal reasons why research in the social sciences has not progressed as far as that in the biological and physical sciences?

The instructor's answer might be:

Since the social scientist is himself part of what he is attempting to study, he cannot achieve the objectivity possible in the more precise sciences. Further, the conclusions he reaches frequently run counter to deeply held prejudices of people, hence are unacceptable. Feeling that many of the social affairs of men are not susceptible to scientific study, people have been less willing to subsidize social research than in medicine, for example. Lastly, the scientific study of nature has a much longer history than the scientific study of man. This history has provided a much larger body of data and theory from which to move forward.

After writing his answer, the teacher might extract the essentials and assign quantitative weights to each. To these he might add items of organization, clarity of expression, and a weight to be used if students present ideas which he had not thought of. This score sheet might look something like this:

1. Scientist part of this subject..... 1
2. Prejudice 1
3. Lack of financial support 1
4. Short history 1
5. Small body of facts and theory..... 1
6. Organization..... 1
7. Language usage..... 1
8. Additional correct information..... 1
9. Incorrect statements . . . . . -1 each statement

Even with such care in preparing a key to an essay

response, subjective factors are likely to creep in. This tends to invalidate the test as an adequate basis for the appraisal of student ability. Many of these factors can be avoided by further refinement of the scoring process.

For long responses and for complete answers, it is advisable to make a gross sorting of responses first — into about three levels of excellence. Then, if finer distinctions are desired, make sorts within each group. It does not seem feasible to use this technique for relatively long tests composed of short answers.

Adapted from Testing Bulletin No. 2, Board of Examiners, Michigan State University, September, 1956.

IV. Grading Issues

While there is no single approach to the assignment of course grades, there are issues to consider in making decisions about grading policy.

1. What are the components of a course grade?
2. Should norm referenced or criterion referenced grading be used?
3. Should grades be curved, based on standard scores, or based on an absolute standard?
4. Are grading policies clearly stated in the course syllabus?
5. Are grading criteria flexible?
6. What is the policy for make up examinations, late papers, dropping a grade?
7. Should you penalize students for guessing on an examination?
8. How should course grades be calculated?

The grades assigned to students in a course may be quite different from different instructors. Listed here are objective test scores for several students in a course. Grades for two students — Nancy and Jeff are compared.

Student	Test #1	Test #2	Test #3	Total	Ave. Score	Discussion:
Jeremy	50	51	30	131	43.6	These test scores
Pat	52	51	40	143	47.6	are not uniformly
Lester	51	50	65	166	55.3	variable from
Sue	53	50	65	168	56.0	test to test. The
Bill	55	51	85	191	63.6	first two tests
Carol	54	51	85	190	63.3	have a homogeneous
Mona	56	50	91	197	65.6	distribution of
Jeff	65	61	63	189	63.0	scores with little variability, and Test 3 scores are quite variable.

By describing several common grading practices, it is possible to observe how Mona and Jeff can receive different grades for the same work from different professors.

Professor	Grading System	Sample Grade Computation
A	Assign grades for each test based on ranking of students; average test grades to arrive at a final grade.	*Mona:* B C A = 3.0 = B *Jeff:* A A C = 3.3 = B+
B	Determine total points of all test scores and assign grades by student rank in class.	*Mona:* 1st, 197 pts. = A *Jeff:* 4th, 189 pts. = C
C	Calculate average of test scores: assign grades by rank in class.	*Mona:* 1st, 65.6 = A *Jeff:* 4th, 63.0 = C
D	Calculate average of test scores; assign grades by clustering of scores.	*Mona:* 65.6 (markedly different from next cluster of 63-64), = A *Jeff:* 63 (within 63-64 cluster) = B
E	Calculate a standard score for each test, then average standard scores. Assign grade based on determined criteria (number of standard deviations from mean).	*Mona:* (.34) + (-.53) + (1.14) = .95 + .95-3 = + .31 = C *Jeff:* (2.39) + (2.62) + (-.11) = 4.9 + 4.9 - 3 = + 1.63 = A

A standard score represents the number of standard deviations a score is from the mean score of a group. It is calculated as follows:

$$Z = \frac{(X - M)}{S}$$

where: Z = standard score
X = raw score
M = mean raw score
S = standard deviation of raw scores

Although all five procedures for assigning grades may be legitimate in most cases, in this particular example Professor E's grades reflect Jeff's overall accomplishment. The grades below were based on the distribution of average Z scores assuming each test was equally weighted.

A = +1.5 to +2.5 SD of M;

B = +.5 to +1.5 SD of M;

C = -.5 to +.5 SD of M;

D = -.5 to -1.5 SD of M;

F = -1.5 to -2.5 SD of M.

Some faculty use a criterion referenced grading system in which absolute standards are set for each grade at the beginning of the course. A common standard is to use 90 percent correct for a grade of "A", 80 - 89 percent correct for a "B", etc. This approach can work well if the standards reflect real differences in student achievement. In some cases the score distributions might reveal that student scores clustered in several groups, but a cluster might begin at 89 percent and end at 95 percent. In this case an arbitrary standard of 90 percent penalizes some students. It is generally a good practice to allow some flexibility in the grading system that allows for unusual situations. Some instructors use the testing system to provide flexibility by allowing students to drop a lowest score, take an optional exam, do a special assignment, etc. to reduce the need for consideration of individual student problems in meeting course requirements. Other instructors have grading policies that consider improvement, motivation, etc. in the final assignment of course grades. Regardless of the grading method used, it is necessary to inform students in the course syllabus about grading policies and to apply them consistently.

V. Test Administration

This section of the handbook, gives procedures designed to contribute to a smooth and uniform test administration. It is primarily intended for test administration in large classes, though some procedures may be useful in classes with 20 or 30 students. The discussion will include the following topics:

- preparing alternate forms and use of form codes
- numbering test books
- seating students
- passing out materials
- collecting and accounting for materials
- cheating
- additional suggestions

■ *Preparing alternate forms and use of form codes*

It may be desirable to create alternate forms of a test for two reasons: to prepare equivalent forms of a test to use at different times and to control for cheating by rearranging items on an alternate form. One method to use when preparing equivalent forms is to select a common set of items to use in each of the equivalent forms. These items are called anchor items. The number of anchor items to use is variable. Selecting approximately 50% anchor items would be a conservative approach. The average difficulty of the anchor set can be compared across forms, and the difference in the ability of two groups of students can be estimated.

Rearranging identical items on a second form of a test is a popular method to control for cheating. The idea is that a student cannot cheat effectively by looking at his neighbor's testing materials because the items are in different orders. Items can be rearranged from "back-to-front", "inside-out" (moving items that are in the middle of one form to the front or back of a second form), or any other pattern that serves the purpose.

Form codes are used to label the alternate forms. For example, two forms of an equivalent test could be labeled A and B. If a scannable answer sheet is to be used, be sure that the form code can be recorded on the answer sheet. Some instructors pre-mark the form code on the answer sheet and insert the answer sheet in the test. The tests are spiralled (form A, form B, form C) and then numbered.

There are some expedient ways to prepare alternate forms. When preparing an exam, prepare three master copies of the test without item numbers or page numbers. Then rearrange the sections of two of the forms and number each form. Once the forms are completed, then xerox the required number of each form. If you are preparing two different forms of a test that have two pages each, a page from each form can be substituted into the other form to create four forms of the test.

Some instructors make alternate forms by keeping the questions in the same position on each form but rearrange the

order of the responses. This has the obvious disadvantage of losing control of the correct response for an item. There is a computer generated test package for the NERDC mainframe that CIRCA maintains (PARKIN01). It uses a random number generator to select questions and rearrange responses.

■ *Numbering tests*

Faculty may want to use numbered tests to provide test security. Numbering tests allows them to be accounted for, which is important if tests are to be reused. Some instructors require students to record their test number on the answer sheet and/or sign the test to be certain every test is returned. These numbers can be used to post student grades.

■ *Seating students*

When students are admitted to the test room, they may be assigned seats by a person supervising the seating inside the test room. This seating of students controls for cheating. The following guidelines can be observed:

1. Students may be assigned to a specific row or column or to a specific seat. Students are not to select their own seats.

2. Students who are obviously acquainted should be separated.

3. All designated seats in each row or column should be filled to expedite the distribution and collection of test materials.

4. If possible, left-handed students are to be seated in chairs with left-handed tablet arms. If use of chairs with right-handed tablet arms cannot be avoided, left-handed students are to be seated with vacant chairs to their left for use as writing surfaces.

5. If at all possible, students are to be seated so that the test supervisor and proctors have unimpeded access to every student.

6. If space permits, use only every other seat to provide greater distance between students. A three-foot separation is desirable.

■ *Passing out materials*

Specific procedures are used in standardized testing when distributing test materials. While these procedures are tedious and time consuming, they are necessary in order to ensure test security. The following general conditions help expedite the process:

1. Test books are distributed across rows or down columns in serial number order. With multiple forms of the test, this distribution decreases the likelihood that adjacent examinees have the identical test form.

2. Where unimpeded access to each student is possible, a test book should be handed to each student, one at a time, in serial number order, along a row or column.

3. If materials must be passed along rows or columns, the exact number of test books is to be counted to supply one book per student. As the materials are passed, each student is to take only one test book from the top of the stack.

■ *Collecting and accounting for materials*

Tests should be collected in the same manner they were distributed. If materials were passed along rows or columns, the tests should be counted as they are collected from each row or column to ensure that the exact numbers are returned. If tests were handed in serial number order to each student, they should be collected one at a time in numerical order.

To account for tests during an administration, there are several obvious suggestions:

- Tests should never be left unattended.
- During the testing session, the number of tests in use along with those not in use should be checked to be sure all tests are accounted for.
- At the end of the testing sessions an inventory of the tests should be made immediately after they are collected.
- If the students are allowed to leave the room as they finish the examination, a proctor should be posted at each exit to collect test materials.

■ *Cheating*

The following information about cheating is extrapolated from the Academic Honesty Guidelines and the "Faculty Guide to Academic Honesty" prepared by the Office of Student Judicial Affairs, 124 Tigert Hall, 392-1261.

Cheating is defined as the giving or taking of any information or material on academic work considered in the determination of a course grade. Taking of information includes, but is not limited to, copying graded homework assignments from another student; working together with another individual(s) on a take-home test or homework when not specifically permitted by the teacher; looking or attempting to look at another student's paper during an examination; looking or attempting to look at text or notes during an examination when not permitted. Tendering of information includes, but is not limited to, giving your work to another student to be used or copied; giving someone answers to exam questions either when the exam is being given or after having taken an exam; informing another person of questions that appear or have appeared on an exam in the same academic term; giving or selling a term paper or other written materials to another student.

To help prevent cheating; faculty may take some of the following steps:

- Discuss academic honesty and the consequences of academic dishonesty with students at the beginning of the term.
- During an exam, seat students in every other seat, when possible.
- Use multiple forms of a test, so students seated next to each other have different forms.

- Number and collect tests, so all tests are accounted for.
- Use proctors to assist in large classes. The Student Honor Court (392-1631) will provide proctors free of charge if there is adequate advance notice.
- Check students' identification, and require students to sign their tests. One very large course with multiple sections requires its teaching assistants to proctor its exams. Then each student is required to hand the completed answer sheet to his or her T.A., who can personally identify the student.
- Handle draft copies of the test as secure material. Be Careful when disposing of draft copies, and do not allow undergraduate students to type or duplicate tests.

How should a faculty member handle an incident of cheating or suspected cheating? Any confirmed incident of cheating must be observed by the instructor and should result in the student's dismissal from the test room. The instructor should attach a report of the incident to the test materials. If the instructor suspects a student is cheating, but the misconduct has not been confirmed beyond a reasonable doubt, the examinee should be warned that cheating is suspected.

If the student is suspected of copying from another student, the instructor may move the suspected student to another seat and should record the name(s) of person(s) from whom the student may have been copying. Then the responses to the questions can be compared later.

If a proctor suspects a student of cheating, he/she should immediately report the incident to the instructor. The instructor should then position himself/herself to observe the suspected student.

After the incident, the faculty member should contact the Director of Student Judicial Affairs (392-1261, 124 Tigert Hall) to determine if it is a first offense. If it is a first offense, several options are available.

1. The student and faculty member may reach an agreement (reduced or failing grade for assignment or course) and sign a Faculty Adjudication Form (available in 124 Tigert Hall).
2. If the faculty member and student do not agree on the matter, they may decide to use the Faculty Determination Process (informal hearing with Director of Student Judicial Affairs, student, student's advisor, faculty member and witnesses) or have the case referred to the Student Honor Court (formal hearing).

If the incident is the student's second offense, the case should be sent to the Student Honor Court.

This discussion has focused on cheating as it relates to test administration. Academic dishonesty also includes plagiarism, bribery, conspiracy and misrepresentation. For more information, faculty may contact the Assistant Dean for Student Services and Director of Student Judicial Affairs (124 Tigert Hall, 392-1216) or refer to the brochure "Faculty Guide to Academic Honnnesty," available in 124 Tigert Hall.

■ *Providing Feedback on Examinations*

There are many different ways that instructors review test results with students. In one case, new examination forms are prepared for each test, and students may keep their tests (while turning in machine-scorable answer sheets). After the testing time has expired, all remaining students are dismissed from the room. Then the room is

cleared, and the correct answers for all forms are displayed on overhead projectors. Students who choose to return to the testing room may score their own exams. This immediate feedback is very popular with and useful to the students.

Other faculty members review the item analysis to identify particularly difficult items. These items are put on overhead transparencies and reviewed at the next class session. This approach is useful when exams are turned in, and answer sheets are not returned to students because of the possibility of changing a response.

Another approach often used to review exams is to hold special test review sessions that are voluntary for students. These sessions may be held in office hours or scheduled on a given date. Copies of the test are distributed for review and then collected.

VI. University Testing Policies

■ *During Term Assembly Examination Policy*

As a reminder, specific examination periods have been established for examinations offered during the academic term. This policy, as approved by the Council of Deans, has been in place since Fall 1982. Please note that the examination times reflect any approved changes to the classroom schedule.

Monday, Tuesday, Wednesday, Thursday, and Friday (M,T,W,R,F) from 5:10-7:10 p.m. (10-11 periods).

If classes are scheduled during examination time, instructors must provide make-up class work for students who miss class because of an assembly exam.

Scheduling for examination periods is to be done through the college scheduling coordinator and priority for examination periods will be determined by the coordinator. College scheduling of two examinations during the same examination period will not be permitted except in extenuating circumstances. If two must be scheduled at the same time, the college coordinator will determine which course should be responsible for makeups. **If a conflict arises between two or more colleges, the course with the larger number of sections will have priority.** If it is not possible to resolve a conflict arising between two or more colleges, the matter should be referred to the Office of Academic Affairs.

■ *Dead Week Policy*

(1) No final examinations should be given during Dead Week except for laboratory examinations.

(2) No assignment of papers or projects shall be made during Dead Week. A take-home final examination can be distributed during Dead Week but cannot be due before the regularly scheduled final examination.

(3) Written papers and/or oral presentations which are announced in the course syllabus distributed at the first class meeting can be collected or presented during Dead Week but this shall not be used to circumvent the intent of Dead Week and (1) above by serving as a final examination. If there is no final examination, papers assigned to replace a final examination are to be due at the time of the scheduled final examination.

(4) Modular quizzes which are taken at a time selected by the student may be administered during Dead Week.

(5) Weekly (or daily) tests which are scheduled in the syllabus distributed at the beginning of the semester are permitted during Dead Week.

(6) Third or fourth hourly tests should not be given during Dead Week.

(7) All changes in the published examination schedule must be approved by the Office of Academic Affairs.

The method of determining grades is the prerogative of the classroom instructor and there is no necessity for a final examination if the instructor does not wish to schedule one. However, the intent of the policy is to protect students (and other instructors) from the necessity of studying for a

comprehensive examination administered during the last week of class when the student should be concentrating on classroom work and beginning to review for final examinations.

■ *Posting Student Grades*

The Privacy Act prohibits posting student names or social security numbers to report test grades or course grades. Alternatives to the use of names and social security numbers can be used such as the assignment of a special unique code to each student which is used for posting grades.

References

Bloom, B. S. (Ed.) Taxonomy of educational objectives. Handbook 1. *The Cognitive Domain*. New York: David McKay, 1956.

Carlson, S. *Creative Classroom Testing*. Educational Testing Service. Princeton, New Jersey: 1985.

Carlson, S. *Item Specific for the College Level Academic Skills Test*. Department of Ecudation, Tallahassee,Florida, 1990.

Downing, Steven. *True-false and alternate-choice formats: A review of the research*. Paper presented at the annual meeting of the National Council on Measurement in Education. Boston: April 1990.

Ebel, R. L. *Measuring educational achievement*. Englewood Cliffs, N. J.: Printice-Hall, 1965.

Ebel, R. L. How to write true-false test items, *Educational and Psychological Measurement*, 1971, *31* (2), 417-426.

Editorial Manual for Item and Test Development. Instructional Research and Practice, University of South Florida, 1986.

ETS Builds a Test. *Educational Testing Service*. Princeton, New Jersey, 1965.

Frisbie, David. *The evolution of the multiple true-false*. Paper presented at the annual meeting of the National Council on Measurement in Education. Boston: April 1990.

Gronlund, Norman E. *Stating Behavioral Objectives for Classroom Instruction*. MacMillan Company, New York, 1970.

Hills, G. C. & Woods, C. T. Multiple true-false questions, *Education in Chemistry*, 1974, *11*, 86-7.

Lord, F. M. The relationship of the reliability of multiple choice tests to the distribution of item difficulties, *Psychometrika,* 1952, 18, 181-194.

Multiple-choice questions: A close look at Educational Testing Service. Princeton, New Jersey: 1963.

Roid, G. H., & Haladyna, T. M. *A technology for test-item writing*. New York, NY: Academic Press, 1982.

Thorndike, R. L., & Hagen, Elizabeth. *Measurement and evaluation in psychology and education* (3rd edition). New York: Wiley, 1969.

Ward, William C. *Ill-Structured Problems as Multiple-Choice Items*. Educational Testing Service Research Report, 83-6. March, 1983.

Wesman, A. C. Writing the test item. In R. L. Thorndike, *Educational Measurement*. American Council on Education. Washington D.C.: 1976.

The Writing of Essay Questions. Testing Bulletin No. 2, Board of Examiners.

Wyrick, Jean. *Steps to Writing Well*. Holt, Rinehart and Winston, 4th ed., 1990.