CONCEPTUALIZING, PLANNING, AND CONSTRUCTING CLASSROOM TESTS

Reginald L. Jones

A previous article by the author in this journal (Principles of Measurement and Evaluation) presented general principles of evaluation and measurement of use to classroom teachers. The present article builds upon the first by presenting a dicussion of (1) the merits of subjective versus objectve tests, (2) considerations in planning the classroom test, and (3) principles of constructing objective tests.

SUBJECTIVE VERSUS OBJECTIVE TESTS[1]

The distinction between subjective and objective tests is a simple one. Subjective examinations are ones wherein the grader is required to use his judgment in evaluating the worth of each response. Objective examinations are ones wherein the evaluation of any response is independent of the grader's judgment. The essay examination is, perhaps, the most commonly used form of subjective test, although oral examinations, themes, speeches, term papers, classroom reports and a host of other techniques are similarly classified. Objective tests typically have a true-false, multiple-choice, or matching format.

The relative virtues and disadvantages of these methods of testing are widely discussed by both students and teachers. Those who criticize objective testing do so on the grounds that it does not permit students to clarify their responses, that it measures superficialities, that objective items tend to be ambiguous, and that this type of measurement may be merely a legalized form of gambling wherein the best guesser is awarded the highest grade. Opponents of subjective testing procedures point to the inordinate amount of time consumed by reading and grading essay papers, the fact that the ques-

1 An earlier version of this section was prepared by Dr. Laurence Siegel.

tions are often so all-inclusive as to prohibit comprehensive treatment by the student, and the possibility that the grades may be influenced by extraneous factors such as legibility of handwriting and grader fatigue.

All of the above-noted arguments contain an element of truth under certain circumstances. The simple fact of the matter is that there are advantages and disadvantages inherent in both subjective and objective testing. The disadvantages are maximized when an inadequate examination is developed; and it is equally possible to construct an inadequate examination of either type.

There are a number of factors which should influence the decision about whether to use a subjective or an objective test format, assuming that the test will be well constructed regardless of its format. This section of the present article is devoted to a discussion of these factors.

CLASS SIZE

A primary advantage of objective testing is the relative ease with which such examinations are scored. The factor of simple and rapid scoring becomes particularly important as class size increases. A one-hour objective examination consisting of seventy-five items, for example, might be hand scored at the rate of about fifty papers an hour, or machine-scored at the rate of about 800 papers an hour. This contrasts sharply with the rate at which a one-hour essay examination (consisting of perhaps five questions) can be graded. If the class consists of about 200 students, the amount of time required to score a one-hour examination might vary between 15 minutes (for the objecuive, machine-scored examination) and 100 hours (for the essay examination). Even hand scoring of the objective examination in this situation would require only four hours or so.

The desirability of rapid scoring from the instructor's point of view need not be labored here. There is, however, one advantage of rapid scoring that is sometimes overlooked. It is most desirable, from an instructional standpoint, to discuss an examination in class, thereby making use of it as a learning experience. Such discussion is most meaningful to the student when it follows closely after the administration of the test and when the student is informed regarding the errors he has made. Since careful grading of subjective examinations is time consuming, such examinations are often

discussed in class without the benefit of simultaneous feedback to the student about his errors. When such feedback is attempted, the delay between administration and discussion of the test may diminish the value of the classroom discussion.

The conscientious instructor who has administered an essay examination to a class of even moderate size is thus confronted by a dilemma. He may read the papers carefully and diminish the value of the test as a learning experience for the student; or he may read the papers more rapidly in order to accelerate the feedback process at a sacrifice of reliability.

RELIABILITY OF SCORES

"Reliability" refers to the consistency of measurement. A reliable test yields relatively stable scores. Although a number of factors may influence the reliability of a test, two are particularly germane to a discussion of the comparative merits of subjective and objective measurement: inter-scorer agreement and intra-scorer agreement. The first of these reflects the extent to which the test score is influenced by variations within a particular grader as a function of such factors as mood fluctuation, fatigue and changing standards as to the constituency of a "good" or "poor" response. Intra-scorer agreement refers to the stability of test scores assigned by different graders. Objectively scored examinations are considerably more reliable than subjectively scored examinations on both counts.

SAMPLING AREAS OF KNOWLEDGE

The fact that each objective item requires relatively little time for response and scoring makes it possible to construct objective examinations which sample a relatively wide band of knowledge. The usual one-hour multiple-choice examination, for example, contains between fifty and seventy-five items; the typical one-hour essay examination may contain only four to six questions. Since the coverage by a subjective test is necessarily limited by considerations of time, a student's performance on it may give an untrue picture of his breadth of knowledge.

An additional strength of the objective examination, related to the fact that it generally contains a more extensive sampling of areas of knowledge than does the subjective examination,

is its particular suitability for diagnostic purposes. A diagnostic analysis of the responses may help the individual student to discover the areas in which he is deficient, and may enable the instructor to discover the areas in which a large percentage of students experienced difficulty.

THE FUNCTIONS MEASURED

The primary advantage of subjective testing is that it facilitates evaluation of certain unique functions which do not readily lend themselves to objective measurement. A note of caution regarding the superiority of subjective testing in this regard is appropriate, however. Quite often the use of subjective methods is improperly justified on the grounds that they measure a unique function whereas, in reality, the same function could be objectively measured. Subjective questions are often incorrectly presumed, for example, to be superior to objective questions when forcing the student to integrate isolated bits of information.

Two unique kinds of functions can, however, best be measured by subjective tests. The first of these is effectiveness of expression. When the instructor is interested in evaluating not only the knowledge possessed by the student, but also his ability to organize and communicate this knowledge in the absence of clues, subjective testing is the method of choice. Secondly, when the instructor wishes to follow sequence of thinking leading to a particular conclusion (e.g., steps in a mathematical derivation; or arguments supporting a particular case solution in business law) he may often do so more adequately by administering a subjective rather than an objective tests. Information about certain educational objectives and their measurement presented in an earlier bulletin, however, suggests that it certainly is possible to measure the aforementioned objectives using objective tests but that much care will have to go into their construction.

It is impossible to generalize about the relative superiority of subjective or objective testing without knowing something about the situation in which the test is to be administered and the purposes to be served by the test. If the class is large, the relatively lengthy time required to construct good objective items is more than offset by the speed with which such papers can be scored. For very small classes, however, it probably will require even more time to develop a good one-hour objective examination than it will to write four or

five essay questions and evaluate the students' responses. The superiority of objective measurement from the standpoint of reliability is undeniable under any circumstance.

The unique suitability of subjective measurement for the evaluation of certain kinds of functions, on the other hand, may occasionally dictate some sacrifice of reliability in order to gain validity. It is most important to recognize that the specific advantages of either type of measurement may be lost whenever the test is improperly constructed or scored. Consequently, a section of the present Article and a future article will be devoted to a discussion of the principles of construction and techniques for scoring objective and subjective (essay) tests.

PLANNING THE CLASSROOM TEST

The primary focus of the present section of the paper is upon techniques which may be utilized in planning the classroom test. These techniques apply equally well to the short, 10 to 15 minute quiz, and to the hour or two hour examination.

Some Considerations in Planning the classroom Test.

The construction of classroom tests requires a considerable amount of preliminary planning involving a number of decisions which are arbitrary, for the most part. Many of these considerations are obvious. A decision must be made, first, about the areas to be covered in the text. Thus, a first test in psychology may cover the topics "Scientific Method in Psychology," and "Methods of Observation" as treated both in the assigned readings and in classroom discussion. Secondly, the question of coverage must be considered. Is coverage to be general (i. e., measurement of student familiarity with a large number of concepts, facts, and principles) or is it to be specific (i. e., measurement only of information presumed by the instructor to be worthy of assimilation by the student)?

A third decision preliminary to planning the test involves consideration of the objectives to be attained. The instructor has to ask himself, "What am I trying to measure". His response to this question may take him in several directions. Some of the more common objectives of measurement are: knowledge of terminology: knowledge of classifications and

categories; knowledge of methodology within the subject-matter field; the application of principles and generalization to concrete situations; a synthesis of several points of view or positions; extrapolation beyond information presented in the class and textbook. It is somewhat paradoxical that although the latter objectives (i. e., application and extrapolation) are generally the most valid ones, they are often neglected in quizzes and hour examinations. One factor which seems to inhibit the writing of questions which measures these more complex thought processes is the fact that such items are relatively difficult to phrase. It is much easier, for example, to construct the item:

Jean Valjean was first sentenced to the galleys for stealing:

a. the Bishop's candlesticks
b. a loaf of bread
c. a few sticks of wood
d. a widow's cow
e. the cloth from the altar

than to construct the item:

Galileo investigated the problem of the acceleration of falling bodies by rolling balls down very smooth planes inclined at increasing angles. He used this procedure because he had no means of measuring very short intervals of time. He extrapolated the case of the free fall from the data he obtained. Which of the following was an assumption implicit in this extrapolation?

a. air resistance is negligible in free fall.
b. objects fall with constant acceleration.
c. the acceleration observed with the inclined plane is the same as that involved in free fall.
d. the planes are frictionless.
e. a vertical plane and one which is nearly so have nearly the same effect on the ball.

However, it is possible to measure the higher order educational objectives using objective test questions although their construction will take some care. The details of multiple-choice item construction will be covered in the section following.

The scheme presented in the following sections is designed to provide an efficient and convenient form for planning the

classroom test. It spotlights the course objectives and content areas and is useful regardless of whether the test to be developed utilizes an objective or subjective format. The test blueprint has several advantages: it assures that the test covers those areas which are presumed to be the most relevant; it is an aid to more efficient test construction; and it permits a diagnostic assessment of sources of student difficulty once the test has been administered.

THE TEST OUTLINE

The test outline is a device for specifying the content of the test. It is predicated upon the instructor's decisions about the content areas to be covered, their relative importance, and the objectives of measurement.

The procedure for developing a test outline may be more efficiently presented by referring to a hypothetical course in educational psychology. Let us assume that lectures and readings have covered the following three topics:

(1) Intelligence, (2) Learning, and (3) Adjustment. The instructor has defined his objectives for this unit as:

1. The ability to recall factual material and principles (specific facts).
2. The ability to apply knowledge and understanding in solving new problems (application).
3. The ability to evaluate unfamiliar situations, ideas or methods in the light of the facts and principles in these areas (evaluation).

The objectives and content areas for this psyhology course are articulated in the test outline as shown in *Figure* 1.

Figure 1. - The Test Outline

Content Area Measured by the Test	OBJECTIVE MEASURED BY TEST: Recall of Specific Fact	Application (60%)	Evaluation (10%)	Total Weighting of Each Content Area
Intelligence	10%	20%	03%	33%
Adjustment	10%	20%	03%	33%
Learning	10%	20%	03%	33%

This outline is structured so that the applicational problems are weighted 60 percent on the examination, whereas factual and evaluation questions are weighted 30 and 10 percent, respectively. Decisions relative to weighting are purely arbitrary and reflect only the emphasis which the instructor wishes to place on each objective and content area.

Writing the Test Items

The construction of text questions is guided in its entirety by the test outline. If we decide that a 51-item objective examination would satisfy the demands of an hour's test, our test outline may take the form shown in *Figure* 2. The entries in our cells now become the number of test items to be included in that specific sub-section.

Figure 2 - Test Outline for a 51-Item Objective Test

OBJECTIVE MEASURED BY TEST

Content Area Measured by the Test	Recall of Specific Fact (30%)	Application (60%)	Evaluation (60%)	Total % for Each Content Area
Intelligence	5	10	2	17
Adjustment	5	10	2	17
Learning	5	10	2	17

The scheme shown in *Figure* 2 lends itself as easily to the construction of subjective (essay) test questions as to the construction of objective test items. We may prefer to write a single essay question, or a problem to replace the 10 objective application items in the adjustment area. Decisions of this type are based upon factors other than the test outline. The outline helps us decide *what* to measure. It does not tell us *how* to measure it.

Evaluating the Test Through Use of the Test Outline

Once the test has been administered, student responses can be appraised to yield some feedback relative to the effectiveness of the teaching and test construction efforts. An instructor may wish to know how well the information within each of the content areas has been assimilated by his students

and how well the objectives within each of the content has been mastered. While answers to those questions will depend, in part, upon the results of item analysis, the test outline can also provide some of this information. Notation of the items posing the greatest difficulty for students may yield clues about areas which need further clarification. The circled items in *Figure* 3 represent those items missed by more than one-half of the students to whom the hypothetical educational psychology test was administered. It is apparent that the application of information relative to "intelligence" seems to be the greatest source of student difficulty. This kind of analysis can provide useful clues serving to aid the instructor as he modifies his course.

Figure 3 - Test Outline for Evaluationg Sources of Student Difficulty

OBJECTIVE MEASURED BY THE TEST

Content Area Measured by the Test	Recall of Specific Fact	Application	Evaluation
Intelligence	1,[2],3,13,14	[16], [17], [27],28 29,30,[31], [37] [38],45	18,24
Adjustment	4,5,6,7,8,	19,20,21,22 46,26,[32],36 39	25,43
Learning	9,10,[11],12	etc.	23

Using the Test Outline to Evaluate Previous Tests

A test outline may also be used to evaluate tests previously given. The purpose of such a procedure would be to appraise the coverage of tests developed earlier in the year without the benefit of a preliminary outline.

Figure 4 - Test Outline for Evaluating Previous Tests

OBJECTIVE MEASURED BY THE TEST

Content Area		Recall of Specific Fact	Application	Evaluation
	Percent Specified	30	60	10
Intelligence	33	1,2,3,14 15,18,19 20,39 9 - 18%	27,35,36,48 4 - 8%	29 1 - 2%
Adjustment	33	5,6,7,12 13,16,21 25,26,40 10 - 20%	28,30,31,32 44,45,46,47 8 - 16%	50 1 - 2%
Learning	33	4,8,9,10 11,17,22 23,24,41 42 11 - 22%	33,34,37,38 43,49 6 - 12%	0%
Total	99	60	36	4

Let us suppose that the educational psychology test which we have been discussing was constructed without an outline. By making a chart similar to the one in Figure 4, the structure of the test may be compared with the outline subsequently developed. Each of the test items is classified according to its content area and objective. The percentage of items within each cell is then compared with the percentage specified by the test outline.

The analysis summarized in Figure 4 suggests that the test which was administered was quite unbalanced. The most significant trend indicated by the test outline is the disproportionately heavy weighting of items calling for factual recall. The test outline requires that only 30 percent of the items be of this type whereas in reality 60% of the test items fell

within this classification. This excess of factual items is reflected in the concurrent underweighting of items in the application and evaluation areas. More specifically, the analysis would indicate to the instructor that his grades assigned on the basis of this test were unduly influenced by a factor (factual recall) which he considers to be of minimal importance and did not reflect achievement of the higher order objectives upon which he wished to place greater emphasis.

This section has presented a scheme for preparing the classroom test. Use of the test outline may seem cumbersome. In its defense, however, it can be asserted that the test outline provides an excellent base for the construction of tests which are properly weighted with respect to content coverage and instructional objectives. In addition, the outline's provisions for content analysis provide a method whereby the instructor may, himself, evaluate the success of his teaching and question writing techniques.

CONSTRUCTING OBJECTIVE TESTS

This section is organized into two major parts. A discussion of certain general principles applicable to the construction of objective tests is presented in the first part while specific principles to be considered when phrasing multiple-choice, short answer, matching, and true-false items are treated in the second.

The construction of the objective test item is a task requiring some skill. Among the requisites of the successful item writer are a thorough knowledge of the subject matter; an intimate understanding of the specific teaching objectives; an insight into the abilities, backgrounds and, particularly the mental processes of the students who are to take the test; a facility in the clear and economical use of language; and perhaps above all, a willingness to devote the time and energy necessary to the task. The instructor preparing the objective test item does not have to be a Shakespeare or a Browning, say, who is adept at creative writing of the sort required for production of a play or poem. However, he-she must have skill in expository writing. In constructing test items the instructor does not need to sit idly while awaiting inspiration for test items. Rather, using the test blue-print as a guide, he may actively stimulate ideas by reference to the lecture notes, text assignments, course outlines and other sources connected with the course.

GENERAL SUGGESTIONS FOR CONSTRUCTING OBJECTIVE ITEMS

The ideal objective test item is phrased in such a way that only those students who possess the required information will make the correct responses. Such items must not contain any internal cues to aid the student who does not possess the necessary information. Questions that can be answered on the basis of general knowledge are avoided. The central problem is clear and each item is independent of every other item in the test. Unessential specificity and irrelevant sources of difficulty are avoided. Each of these principles is discussed and illustrated below.

A. Avoid Internal Cues

Example I

Internal Cues Eliminated

The probability equation can be used to determine the likelihood of a student answering an objective item correctly on the basis of a lucky guess.

Example II
Internal Cues Provided

Because the T-F test contains only two response alternatives, whereas the multiple-choice item contains four or more alternatives, the probability of a lucky guess is much higher for T-F items than for multiple-choice items.

The correct answer for both items is "true." The internal cue provided by the phraseology of Example II gives away the answer. Example I, however, requires the student to arrive at the correct generalization and then to check this generalization against the problem posed by the test item.

Internal cues are sometimes provided also by the appearance of common elements in the lead and the correct alternative of multiple-choice items. An obvious illustration of this kind of defect is provided by the following item:

Item: A resolution of the epistemological dualism which emphasizes *mentalistic* phenomena is known as:

a. *mentalistic monism*
b. materialistic monism
c. parallelism

A final source of internal cues is the appearance of interrelated items in the test. Information provided in one item may, if care is not exerted in test construction, provide the alert student with the answer to some other item in the test.

B. Avoid Questions That Can Be Answered on the Basis of General Knowledge:

Items answerable solely on the basis of general knowledge do not reveal anything about what a student has learned in a particular course.

Unfortunately, the person who has constructed the test may have some difficulty in differentiating between items measuring course-related knowledge and items measuring the student's general background. It may be a good idea, for this reason, to have a person not expertly familiar with the content area read through the test with a view toward the identification of non-functional items.

Example I
Requires General Knowledge

Which of the following test formats would entail the greatest amount of time for grading (assuming that testing time is constant)?

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

Example II
Requires Some Specific Knowledge

The factor contributing *most* to the relatively lengthy time required to arrive at reliable scores for subjective tests is that-

a. each question must be read at least 2 times.
b. problems of penmanship often make the reading of some papers difficult.
c. each paper must be read and compared with every other paper.

d. many rest pauses are necessary during the grading process in order to reduce grader-fatigue.

Since it is common knowledge that essay tests require more time for grading than do subjective tests, almost any adult could answer Example I correctly. The correct response to Example II, on the other hand, requires either that the student has had experience with the problem posed in the item, or that he has been exposed to related didactic material.

C. Clarifying the Central Probelm

Every objective item should contribute to the efficiency of the test in differentiating between those students who possess the requisite knowledge and those who do not. Thus, the difficulty of the item should be a function of the knowledge required by it, rather than of the language in which it is expressed. (The exceptions to this generalization are tests of reading or intelligence wherein the object of the test item may be to measure vocabulary or the ability to understand complex sentences.)

Examples (T-F Items) Poor	**Better**
Students who are obfuscated by complex and pedantic item language often display cognitive response blocks, in spite of prior acquisition of the requisite information.	Unnecessarily complex vocabulary may cause many students to answer incorrectly examination items for which they possess the relevant information.

D. Avoid Unessential Specificity and Irrelevant Sources of Difficulty

Example I
Unessential Specificity

Given the following information about test scores

$N = 627$

$\Sigma X = 38217.441$

$\Sigma X^2 = 172839.456$

The average score, corrected to two decimals, is _____.

Example II
Specificity Minimized

Given the following information about test scores

$N = 10$
$X = 35$
$X^2 = 70$

The average score, corrected to two decimal places is .....

The greater amount of arithmetic manipulation required for solution of the problem in Example I makes no particular contribution to the validity of the item. The purpose of the item is to determine whether or not the student understands how to compute a mean or arithmetical average. This purpose is satisfied by Example II. The specificity of Example I would be justifed if we were interested also in the student's understanding of rudimentary arithmetic processes. Unnecessary specificity increases the amount of time required by each item and thereby reduces the breadth of coverage of the total test.

SPECIFIC PRINCIPLES FOR CONSTRUCTING MULTIPLE-CHOICE ITEMS

The advantages of the multiple-choice format include ease of scoring and the reducction of chance or guessing factors. This format is exceedingly flexible; items may be based upon charts, graphs, or verbal materials.

A. Statement of the Problem

The lead of the multiple-choice item should present a single problem stated with sufficient clarity and precision to permit the student to anticipate the nature of the required response.

Example I - (Good)

The consistency with which a test measures is termed

a. validity
b. reliability
c. standardization
d. homoscedasticity

Example II - (Poor)

Reliability

a. may vary between - 1.00 and 98
b refers to the consistency of measurement.
c. is an unimportant characteristic of a test.
d. indicates the predictive efficiency of the test.

B. Plausibility of the Alternatives

All of the alternatives must be plausible and attractuve to students who do not possess the required information. The distracters (incorrect alternatives) should, if possible, be based upon the usual sources of student error or difficulty. If an item contains one or two distracters which are obviously wrong, the possibility of a student responding correctly solely on the basis of a guess is increased greatly.

Example I
Alternatives Plausible

In planning the classroom test, the instructor must first make a decision about

a. the types of items to be included.
b. the level of coverage he desires.
c. the objectives to be measured.
d. the number of items to be included in the test.

Example II
Alternatives Not Plausible

In planning the classroom test, the instructor must first make a decision about

a. the working area which is most conductive to productive thought.
b. the design of the work-sheet.
c. the objectives to be measured.
d. the level of coverage he desrires.

Virtually all students will discard alternatives *a* and *b* from Example II. The guessing factor is thus increased to 50 percent and the item now assumes the characteristics of a true-false format.

C. Length of the Alternatives

No one alternative within a given test item should be overly long or overly short. The alternative that is atypical in length is generally the correct one. This cue can be avoided either by making the alternatives approximately equal in

length, or by structuring them so that they are all unequal in length.

Example I
Alternative of Unequal Length

Long alternatives in the multiple-choice item are frequently correct because of

a. their greater validity.
b. the greater numbre of qualifying phrases necessary to make them correct.
c. their greater reliability.
d. their greater objectivity.

Example II
Alternatives Equal in Length

Multiple-choice items with alternatives of approximately equal length reduce guessing because they

a. increase item reliability.
b. increase item validity.
c. reduce internal item cues.
d. reduce item subjectivity.

D. Alternatives Phrased in Technical Jargon

The presence of one highly technical distracter provides an internal cue to the alert student. This effect is equivalent to reducing the number of alternatives. Any subject area will include information which can be communicated in varying degrees of technical language. A test item works most effectively, however, when the level of technicality within the item is relatively constant—i.e., all of the alternatives are phrased in technical or non-technical language.

Example I
Technical Jargon for One Alternative

The concept of item stability in relation to factor pattern involves the

a. correlation confficient
b. Wherry-Gaylord procedure
c. reliability coefficient
d. validity coefficient

Example II
No Technical Jargon

The presence of one distracter phrased in technical jargon reduces the

a. item's validity
b. number of effective alternatives
c. item's reliability
d. inter-item stability

E. Partial Parallelism:

Partially parallel alternatives weaken item structure. Alternatives are said to be parallel when their phraseology is identical with the exception of one or two words. This is a defect only when parallelism exists for two of the alternatives and not for the others. The correct alternative, in such instances, is generally found in the parallel pair.

Example I
Partial Parallelism: Converse

When a test is increased in length

a. reliability is increased.
b. reliability is decreased.
c. validity is decreased.
d. inter-item consistency is increased.

Example II
Partial Parallelism: Not Converse

The essay examination would be classified as a (an)

a. subjective test.
b. objective test.
c. time consuming test.
d. unreliable test.

Both of these illustrations contain partially parallel alternatives. The student's attention is narrowed to two choices (*a* and *b*), since both alternatives generally cannot be incorrect. If one of these alternatives is known to be incorrect, the problem posed in the item is automatically solved.

Partially parallel alternatives may be avoided by structuring alternatives all of which are parallel to one nother, none of which are parallel, or by structuring pairs of parallel alternatives.

Example I - Complete Parallelism

When a test is increased in length

a. reliability is variable
b. reliability is decreased
c. reliability is unchanged
d. reliability is increased

Example II - Pairs of Parallel Alternatives

When a test is increased in length

a. the mean is increased
b. the mean is decreased
c. reliability is increased
d. reliability is decreased

Example III - No Parallelism

When a test is increased in length

a. validity is unaffected
b. reliability is increased
c. inter-item variability fluctuates
d. cluster analysis is desirable

F. Unique Alternatives

Unique alternatives like "all of the above" or "none of the above" must be used with care. When used sporadically throughout the test, these alternatives are too frequently the correct answers. These two alternatives may, of course, be added to the items in order to increase the difficulty of the test.

G. Location of the Correct Alternative

Quite often students observe that an instructor tends to locate the correct answer at a given alternative position with a fair degree of consistency. A favorite locational bias in four or five alternative items is position *c*. What probably occurs during the process of test construction is that the instructor constructs his lead and thinks immediately of the

correct answer. He doesn't want to place this alternative first because it appears "too obvious." Consequently, he writes alternatives for the *a* and *b* positions. The third alternative is frequently a difficult one to construct, so he slips in the correct alternative at the c position while thinking of a plausible alternative for the *d* position.

Locational biases may be overcome by routinely drafting items with the correct answer placed as the first alternative. After all items are written, the location of the correct alternatives can be randomized so that they are presented with approximately equal frequency at each of the response positions.

H. The Number of Alternatives

The number of alternatives to be presented for each item is influenced by two considerations: the guessing factor, and practicality. The effect of increasing the number of alternatives is to reduce the possibility of correct responses by guessing. Increasing the number of alternatives, however, increases the difficulty of the task confronting the instructor as he develops the test, and decreases the potential coverage of the test because of the time required to read each item. Items containing four or five alternatives (exclusive of "all of the above" and "none of the above") strike an optimal balance between these factors.

SHORT-ANSWER TESTS

The short-answer format lends itself to problems which involve simple computations or verbal associations. It may take several forms.

1. The question variety

Example: What was the title of Bulletin No. 1 in the HSIU Testing and Grading Series?...

2. The completion variety

Example: Development of the test outline was presented in Bulletin No.....of the HSIU Testing and Grading Series.

3. **The identification or association variety**

Example: Directions: Write the number of the Bulletin in the Testing and Grading Series in which each of the following topics was discussed.

1. Shortcomings of essay tests. .....
2. The test outline. .....
3. Constructing short-answer tests. .....
4. Principles of measurement. .....
5. Item analysis. .....
6. Disadvantages of true-false tests. .....

A principle source of difficulty with short-answer items is ambiguity of response. Since there are so many synonyms and equivalents for concepts, it is quite possible that a given item can be answered correctly in a variety of ways. Thus, the scoring key may be cumbersome and scoring time may be relatively lengthy.

A. Specificity of Response

The short-answer format should be used only for questions that can be answered by a unique phrase or number.

Example I - (Good)

The numerical value of the constant term *e* in the equation for the normal curve, corrected to four decimal places, is.....

Example II - (Poor)

.....and.....are evaluative criteria least characteristic of.....test items.

It is quite obvious that any number of acceptable responses could be made to Example II, whereas the answer to Example I is highly specific and therefore amenable to the short-answer format.

B. Avoid Verbatim Phraseology

Short-answer items based upon verbatim phraseology from assigned readings encourage rote memorization.

C. Construct items that deal with important content; do not measure trivia.

In keeping with specifications in the test blueprint, the information which a student is required to recall on a test should have some importance in the area of study. It is a reasonable question for the instructor to ask, in considering a possible test item, whether knowing or not knowing the answer would make a significant difference in assessment of the students' competence in the area being appraised.

Poor

What is the middle initial of the author of Testing and Grading Bulletin No. 4?.....

D. Make certain that the test question or statement poses a specific problem for the examinee.

The short answer question should be written in such a way that a student who knows the material will know what the desired answer is.

Example I - (Poor)

What is a desirable characteristic of a test?.....

Example II - (Good)

The test characteristic which refers to the consistency of measurement is.....

MATCHING TESTS

This format is particularly suitable for testing factual knowledge (who? what? where? when?) . It is difficult to structure matching items for use in the measurement of understandings.

A. Avoid One-to One Matching.

The use of an equal number of items and matching alternatives provides the student with cues by virtue of a process of elimination. It is far better to structure matching items so that each alternative may be used more than once and some may not be used at all.

B. Homogeneity of Premises and Alternatives

Example I. Homogeneous premises and responses not used.

Directions: Match the lettered alternatives with the appropriate numbered statement,.

1. A procedure that estimates the consistency with which a test item measures.	A. Kuder-Richardson B. .98 C. reliability D. validity
2. Coefficient of correlation representing a high degree of relationship.	
3. Developed a widely used formula for measuring inter-item consistency.	

Example II. Homogeneous premises and responses used. Directions: Match the lettered alternative in Column II with the appropriate type of test in Column I.

Column I - Tests	Column II - Categories
1. Essay	A. Objective test
2. Matching	B. Subjective test
3. Multiple-choice	C. Neither objective nor subjective test
4. Short-answer	
5. True-false	

Internal cues are provided in Example I because it is obvious that item 2 can be satisfied only by response *D*, whereas alternative *A*, is the only one that fits item 3. This kind of cue is avoided when the concepts included in the item are homogeneous (Example II).

C. Relatively short Lists of Alternatives (Five or Six) Should Be Used,

An overly long list of alternatives makes it difficult to maintain homogeneity and increases the probability that a student may make simple clerical errors in marking his answers.

D. Provide Clear Directions:

The directions should explain the intended basis for matching.

Example I - (Good)

Directions Item 1-4 are concerned with the effects of certain changes in test structure upon test reliability. The possible effects with which we are concerned are lettered. These are to be matched to the appropriate alteration in structure, (You may use any alternative more than once; you need not use all of the alternatives.)

1. Addition of 9 homogeneous items to the test.	A. This would reduce the reliability of the test.
2. Addition of 14 heterogeneous items to the test.	B. This would increase the reliability of the test.
3. Addition of 30 homogeneous items to the test.	C. This would have no effect on the reliability of the test.
4. Increase in time limits of the test.	

Example II - (Poor)

A poorer set of directions for the above items would read:
Directions: Match the following actions and consequences.

DICHOTOMOUS RESPONSE TESTS

Dichotomous response tests consist of sentences to which the examinee responds with one of two alternatives, usually right-wrong (R-W), yes-no (Y-N) or true-false (T-F).

The value of this type of test as a measure of educational attainment is a matter of some controversy. Arguments in favor of this format are based upon the fact that such items are easy to construct and are especially appropriate for measuring specific factual material. Brevity and simplicity make it possible to cover a wide range of material. Opponents of the true-false test feel that the format is too ambiguous, measures little except the student's reading ability, and does not lend itself to the measurment of understanding, extrapolation, cause and effect relationships, etc. The scores on this type of test are, of course, highly susceptible to guessing.

A. Structuring True-False Tests

1. The test item must be unambiguously true or false.

Example I Item Ambiguous	Example II Ambiguity Minimized
Studies have shown that true and false items are the most reliable.	A study by Brown showed that the true-false items were scored more reliably than were essay questions.

The student has to guess the degree of tolerance which the scorer will accept when the item is not stated unambiguously. Unfortunately, it is the better students who most often lose credit because of their superior ability to perceive the existence of ambiguity.

2. Avoid the use of long and involved statements with many qualifying phrases. Such statements pose problems for the examinee relative to the identification of the crucial element in the item.

3. Sentences borrowed verbatim from texts or other sources should be avoided. The meaning of these sentences may not be clear when removed from context. In addition, the use of verbatim phraseology encourages rote learning.

4. Avoid the use of specific determiners. Test items that include words such as "all," "never," "no," "always," or other all-inclusive terms represent such broad generalizations that they are likely to be false. Qualified statements such as "usually," "sometimes," "under certain conditions," "may be," etc. are likely to be true. The test-wise student knows this and may use such cues to get credit for knowledge which he does not possess.

5. Avoid the use of negative statements and particularly double negative statements.

Example I - (Poor)	Example II - (Better)
The Taxonomy objective that describes student ability to go beyond the information presented is not called evaluation	The Taxonomy objective that describes student ability to go beyond the information presented is called evaluation.

6. Keep true and false statements approximately equal in length. Quite frequently, on teacher-made tests, there is a tendency for true statements to be longer than false ones. This is so because the true statements need to include qualifications and limitations to make them unequivocally true. However, an occasional long true statement is not serious if it is matched by an occasional long false one, and there is no consistent difference in length between the two categories of statements.

B. Modifications of the True-False Test

1. The correction variety

This modification of the true-false test requires that the student not only make the correct response, but also that he correct any errors in the item. The student is directed to make false statements true by suggesting a substitute for the word underlined.

Example:

Directions: Indicate the correct response by circling the T or F to the left of the statement. If the item is judged to be false, the word underlined should be replaced by one which would make the statement true. This word should be written in the blank to the right of each item.

T-F The essay test is an example of an *objective* examination._____

In this instance, the item would be marked false and the word "subjective" would be inserted in the blank to the right. This type of format obviates machine-scoring, but it does provide a better measure of student achievement than does the T-F test without correction.

2. Qualified True-False Test

The qualified true-false test minimizes the problem of ambiguity in T-F items. Like the true-false test of the correction variety, satisfactory performance with qualified true-false items requires that the student go beyond the information presented in the itme.

Example:

Directions: Each of the following items may be true without qualification, true with qualification, or false. If the item is true without qualification, circle the T and mark a "c" in the space provided. If it is true with one of the listed qualifications, circle the T and mark the letter of the appropriate qualification in the space. If the item is false, circle the F.

	Statements	**Qualifications**
T-F 1.	A longer test is more reliable than a shorter one.	a. If the items are homogeneous.
T-F 2.	A longer test is generally more desirable than a shorter one.	b. If the items are equally valid.
		c. No qualifications.
T-F 3.	The application of the Spearman-Brown formula will yield a valid estimate of the reliability of a lengthened test.	

SUMMARY

Considerations in deciding whether to use objective versus subjective tests, and a description of procedures useful in planning classroom tests were presented in this article. In addition, a number of key principles to be considered in the construction of objective tests were given. However, skill in writing good objective test items comes about only through the actual process of writing items and submitting them to subsequent analysis. It is quite obvious that an item of any type which does not differentiate good from poor students, or which does not assess desired learning outcomes in an unbiased fashion, can have little value for the assessment of student knowledge or for assessing the effect of instruction.