DOCUMENT RESUME

ED 081 784 TM 003 144

TITLE	Making the Classroom Test: A Guide for Teachers.
INSTITUTION	Educational Testing Service, Princeton, N.J.
PUB DATE	73
NOTE	10p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Achievement Tests; American History; Arithmetic; Biology; Elementary Grades; English; Guides; *Scoring; Secondary Grades; Statistical Analysis; *Student Testing; *Teacher Developed Materials; *Test Construction; Tests

ABSTRACT

The plans and procedures used by four teachers in making good tests are presented, and examples are given to serve as guides. The four tests are: An Objective Achievement Test on Fifth-Grade Arithmetic; An Essay Test to Measure a Special Ability in Eighth-Grade American History; A Tenth-Grade Biology Test Especially Designed to Rank Students in Order of Ability; and A Twelfth-Grade English Test for Diagnosing Common Errors in Usage and Spelling. Following descriptions of the tests, basic rules of test-making illustrated by the four tests are given; these are: (1) have the purpose of the test clearly in mind. (2) make a careful plan for the test questions; (3) if test is mainly diagnostic in a basic skill area, prepare at least 10 questions (preferably more) for each subtest used; (4) to find out how well the class has mastered a particular unit of study, make a test which parallels the work in class, and (5) to rank a selected group of students in order of their achievement, the questions hould be on "critical" points of learning. Special problems in writing and scoring tests are determining when essay questions and when objective questions should be used, writing questions that are understandable, stating questions so that there can be only one interpretation, improving reliability in scoring essay questions, the kind of statistical analysis the teacher should make of test questions, and when to use a published test and when the teacher should make his own. (DB)

MAKING THE CLASSROOM TEST:
a Guide for Teachers

EDUCATIONAL TESTING SERVICE · PRINCETON, NEW JERSEY

All teachers have to make tests. But making good tests is not easy. The purpose of this pamphlet is to offer practical suggestions which may help you to make better tests.

To make a good test, you should have clearly in mind what you are testing for and how you plan to use the results. A carefully planned unit of study deserves a carefully planned test covering the unit. As you know, good tests are not made by merely "throwing together" questions more or less related to the work you have been teaching until you have written enough to keep the pupils busy for a class period. A test prepared in a haphazard manner will *not* really tell you how much your pupils have learned. Furthermore, it may well leave your pupils confused about what they are supposed to have learned.

Let us consider the plans and procedures used by *four* teachers in making good tests. Their efforts illustrate basic general principles for constructing tests to meet specific classroom needs. First, the forethoughts and procedures of each teacher will be described. Then the general rules will be stated.

An Objective Achievement Test on Fifth-Grade Arithmetic

It was near the end of the school year and Mrs. Jackson, fifth-grade teacher, decided to give her pupils an arithmetic test covering the year's work. Her first step was to list the kinds of information she hoped to get from the test. She decided that, most of all, she wanted to get a general picture of class achievement with some indication of over-all areas of strength and weakness. Secondary purposes she listed were (1) to identify those pupils who might be especially weak in a particular arithmetic skill and (2) to measure the relative abilities of her students for purposes of report-card grading.

In trying to get an accurate picture of over-all class achievement, she decided that there were two ways in which she could classify the year's work: one was according to *the kind of computation required*, and the other was according to *the way the problem was presented*.

The kinds of computation required were

1. Multiplication
2. Division
3. Addition and Subtraction of Fractions
4. Measuring (distance, time, weight, temperature, etc.)
5. Decimals

The ways the problems were presented were

1. Simple computation, such as $21\overline{)\$1.05}$ or $1/2+1/3$
2. Problems requiring use of procedures learned previously, such as

John missed 1/5 of the twenty words on a spelling test. How many words did he miss?

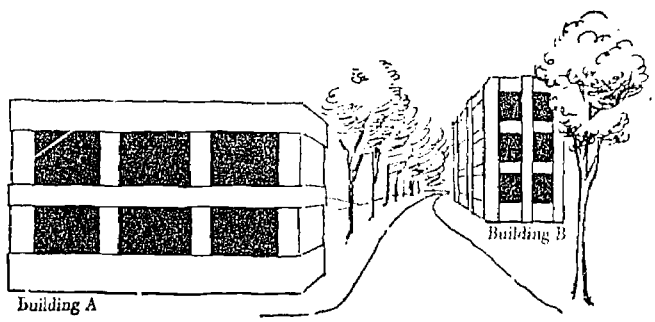
or

A group of twenty-nine children were making programs for a school assembly. They needed 435 programs. How many did each child have to make?

3. Problems requiring original thinking by pupils and use of "number sense." In these problems the pupils could not *depend on previously learned procedures for a method* of solution but must develop their own procedures for solution. Two problems of this type follow:

Problem one. Explain how you, as a fifth-grade pupil, using ten blocks, could prove to a fourth-grade pupil that 1/2 is bigger than 2/5.

Problem two. You are standing directly in front of Building A and looking off at Building B in the distance. Here is the way the two buildings look to you:



The rooms in both buildings are the same height. By looking at the windows, decide which of the following is true:

(A) Both buildings are the same height.
(B) Building A is two-thirds as high as Building B.
(C) Building A is one and one-third times as high as Building B.
(D) Building A is twice as high as Building B.
(E) You can't tell from looking at the buildings which one is higher.

(Answer: B)

Using these two ways for classifying the questions (according to the kind of computation required and according to the way the problem was presented), Mrs. Jackson was now ready to make *a written plan* for the test. She intended that this plan would provide for a test *paralleling* the emphasis given to various points in class.

Mrs. Jackson wrote out her test plan in the form of a "two-way grid." In a two-way grid each question is classified in two dimensions.

The two-way grid that Mrs. Jackson made for the arithmetic test is on page 2. Since she planned to allow an hour for the test, she thought 40 questions would be about the right number. The numbers in the boxes represent ques-

tions--these questions to be of a type described by the two dimensions of the grid.

Kind of Computation Required	*Way Problem Presented*		
	Routine Computation	Thought Problems Following Procedures Taught Previously	Thought Problems Requiring Students to Develop New Procedures
Fractions	7	4	1
Multiplication	2	3	1
Division	2	4	1
Measuring	1	5	1
Decimals	3	3	1

After Mrs. Jackson completed the two-way grid, she found it relatively easy to write most of the questions for the test. She was able to write many questions by paralleling questions from the arithmetic textbook itself. However, she found it quite challenging to write the five problems which would require students to develop new procedures.

Mrs. Jackson believed that the test covered understandings and skills in which her pupils had been well prepared. Therefore, she expected the very best students to get all or nearly all of the questions right, and she expected even the below-average students to get a majority of the questions right. She did not, however, make the mistake of deciding in advance that some minimum score--say 28 questions right (70%)--would represent a passing mark. She knew from previous experience that sometimes her questions turned out to be more difficult than they first seemed to her. She decided to wait until she could look at the scores actually made on the whole test and could scrutinize carefully any questions which proved particularly troublesome.

As it happened, most of the students did well on the test, although no one had a perfect paper. On the basis of the test, Mrs. Jackson felt that her class had achieved the objectives of the work in arithmetic. She did notice, however, that a number of students had difficulty with the problems involving decimals. Therefore she decided to spend more time working on decimals in the few weeks remaining in the school year. And then there was one student who failed all the division problems, although he did fairly well on the rest of the test. She arranged to give this student special help in division.

Most of the students had between 30 and 35 questions right. However, there were a few who scored above, and a few who fell below this middle range. Knowing which students were in the middle and which were above or below was useful to Mrs. Jackson in assigning report-card grades. Of course she also took into account each pupil's class work and his standing on other tests.

In evaluating her test, Mrs. Jackson felt it had been reasonably successful in meeting the purposes for which she had planned it. The test had given her a good picture of over-all class achievement and it had pointed up the weakness in decimals. It had not been planned to be highly diagnostic, but it had helped to identify one pupil who was especially weak in division. In addition, although the test did not rank all of her students in the exact order of their arithmetic abilities, it had given her information that was useful for grading purposes.

An Essay Test to Measure a Special Ability in Eighth-Grade American History

Mr. Frank's eighth-grade American history class had been studying the fighting that took place between the Indians and the settlers in the western states. The class had just completed several discussions on the rights of each side.

The major purpose of having these discussions was to improve the pupils' ability to find and express convincingly facts and arguments in support of their opinions.

Mr. Frank decided that he would like to give a test to measure his class's skill in this ability. At first, he considered giving an objective test. He thought he might list a number of arguments presented by both the Indians and the settlers and then ask the class to identify those which were backed up by facts. But then he decided against using this kind of test. An objective test would require the student to select sound arguments; it would not call upon him to develop and present them convincingly as he would do in actual discussion. Accordingly, Mr. Frank decided that an essay test would satisfy his purposes best.

Since the subject matter of the test was limited, it was unnecessary for Mr. Frank to prepare a written plan for the test. In a sense, the test questions themselves constituted the test plan. Here is the test he prepared. It had three questions:

1. Pretend that you are a settler and give three general reasons why you think your side is right in the war with the Indians. For each of the reasons, describe an actual happening to support your argument.
2. Pretend that you are an Indian and give three reasons why you think your side is right in the war with the settlers. For each of the reasons, describe an actual happening to support your argument.
3. Look at the six reasons given by both sides and decide which one would be most dangerous if everyone accepted this kind of reasoning. Give two examples of how people might do bad things if they accepted this kind of reasoning.

Before scoring the papers, Mr. Frank analyzed the points which he thought would appear in an ideal response and decided how much he would count for each point. He decided not to take off credit for mistakes in spelling and English usage. But he planned to show the English teacher any paper which was especially poorly written so that the English teacher might give help in composition writing to those pupils who needed it.

After Mr. Frank corrected the papers, he found that most of the pupils had proceeded well on Questions 1 and 2 requiring reasons and examples. However, many of them had floundered on Question 3, which required them to point out the dangerous implications of one argument. Because of their difficulty with Question 3, Mr. Frank decided to organize a series of classroom debates, so that the students would get practice in extending, attacking, and defending an argument.

On an essay test of this sort, scores are not highly reliable. On a second reading, after a little time lapse, Mr. Frank would find it difficult to give every paper the same mark as on the first reading. Furthermore, several teachers grading the same papers would probably not agree very closely with one another. Therefore, Mr. Frank avoided giving an exact numerical score for each paper but instead

assigned three general grades: good, average, and poor. However, he wrote many comments on the papers so that the pupils would have a better idea of the strengths and weaknesses of their arguments. He also read several papers to the class for discussion purposes, making full use of the test as an instructional device.

A Tenth-Grade Biology Test Especially Designed to Rank Students in Order of Ability

The students in Mr. Orlando's tenth-grade biology class were all in the college-preparatory curriculum. Moreover, they were all good science students who had been especially selected for accelerated work.

Although it was unlikely that any of these students would fail, Mr. Orlando wanted to find out who were the "A" students, who were the "B" students, and who, if any, should be given "C's." He also wanted to select three students for scholarships to attend a special summer science institute at the state university.

To get a highly reliable ranking of his students on their knowledge and understanding of important topics in biology, Mr. Orlando decided to give a *series* of tests on the various subjects covered in the course. The first test he planned was on a reading assignment about antibodies.

He knew from previous experience in testing this class that, if his questions generally paralleled the entire reading assignment, most of the class would answer them correctly. Such questions would not help him much to rank the students in precise order or to select the three scholarship winners. Therefore, he decided to focus this test on "critical points" in the reading assignment. That is, the test questions would be on the parts of the reading assignment which were hard to grasp and yet were important for complete understanding. These questions would have to be so difficult that average or even good students might well miss them, but the very best students would probably get many of them right. At the same time, Mr. Orlando was very careful to avoid questions which might be difficult simply because they were on trivial details.

He divided the reading assignment into four important content areas. Then, within each of the content areas, he planned to test for two skills which he considered *crucial* for top-flight science students. The four content areas and the two skills within these areas, as well as the number of questions, are presented in the grid below:

Skills: / *Content*	Questions Requiring Exact Definition of Technical Terms	Questions Requiring Application of Information from Reading Assignment
The Antigen-Antibody Reaction	3	10
The Rhesus Blood Factor	3	10
The A, B, and O Blood Types	3	10
The Importance of Antibodies in Disease	1	10

Although Mr. Orlando intended to use objective questions for measuring skill in defining technical terms, he at first thought that he would need to use essay questions for measuring skill in applying information. In fact, he had already written the following question for the test:

You are given two test tubes, one labeled Protein Q, the other labeled Protein Z. How could you tell if these tubes *really contained different proteins*? Outline the experimental procedure you would follow.

However, to answer this question would take many of the students at least ten minutes. It might take some of the better students even longer because they would tend to describe the experiment in greater detail than the others. At this rate, it would take most of the students the better part of the day to finish the test.

Therefore, Mr. Orlando began to consider the possibility of using objective questions instead. He rewrote the essay question, above, in this way:

You are given two test tubes, one labeled Protein Q, the other labeled Protein Z. Which of the following would be the best first step to find out if these tubes really contain different proteins?

(A) Inject a rabbit with either Protein Q or Protein Z.
(B) Mix Proteins Q and Z together to see whether a precipitate is formed.
(C) Take blood from a rabbit and centrifuge out the red blood cells.
(D) Add a serum to either Protein Q or Protein Z.
(E) Inject a rabbit with a combination of Protein Q and Protein Z.

(Answer: A)

To answer this question would require most students only a minute or two. Of course, the objective question above requires simply selection of the *first step* in the experiment, while the essay question would require a *complete description* of it. Yet Mr. Orlando felt that most of those who could select the first step correctly would also be able to carry through the whole experiment. And by using the objective question, Mr. Orlando was now able to ask many more questions than if he had used the essay type.

Here are some other questions he wrote. The first two require an exact understanding of some of the technical vocabulary in the reading assignment. The others require application of information given in the reading assignment.

1. An antigen is
(A) the opposite of an antibody
(B) the residue of an antibody
(C) the stimulus for antibody formation
(D) the result of antibody formation
(E) the same as an antibody

(Answer: C)

2. The antibodies in passive immunity are
(A) created by the recipient
(B) preformed
(C) created by fresh exposure
(D) self-reproducing
(E) created by antitoxins

(Answer: B)

In the following table fill in the possible blood groups of fathers.

Blood Types Determining Possible Parentage of a Child, Making Use of the A and B Proteins Only.

	Known Blood Group of Mother	Blood Group of Child	Possible Blood Groups of Father
3.	A	A	
4.	O	A	
5.	B	AB	
6.	B	A	

(Answers: 3-A, B, AB, O; 4-A, AB; 5-A, AB; 6-A, AB.)

7. A male child is born with erythroblastosis fetalis which requires a blood exchange. Which one of the following must be true?
   (A) The mother was Rh+ and the father Rh−.
   (B) The mother was Rh− and the father was Rh+.
   (C) The mother was B+ and the father was A−.
   (D) The mother had B type blood and the father had A type.
   (E) The mother had O type blood and the father had AB type.
   (Answer: B)

8. Which of the following is NOT an immunological approach to combating disease?
   (A) Salk vaccine
   (B) Gamma globulin
   (C) Penicillin injection
   (D) Tetanus toxoid
   (E) Smallpox vaccination
   (Answer: C)

The class found the test challenging and enjoyed going over the questions with Mr. Orlando afterwards. Although there were a few questions that most of the students got right and a few that practically everyone got wrong, most of the questions were missed by about half the students. The average score in the class was 23 out of the 50 right. The lowest score was 9 right and the top score was 48 right. The test helped Mr. Orlando to rank the students with some confidence. He felt that the test also helped demonstrate to the class the kind of reasoning-from-data required for advanced work in science.

A Twelfth-Grade English Test for Diagnosing Common Errors in Usage and Spelling

It was the beginning of the school year and Miss Barstow had been assigned a special twelfth-grade class in remedial English. All the pupils in her class had done poorly in high-school English. The English Department had decided that these pupils should be drilled intensively to improve their basic skills in composition writing before graduation from high school.

Miss Barstow decided that she needed to use a diagnostic test to find out areas of strength and weakness for each pupil.

The English Department had already designated six basic types of error in composition writing to be corrected in this class. These were:

1. Run-together sentences, or the comma-splice:
   "It is snowing very hard today, the children will probably go sleigh riding tomorrow."
2. Incomplete sentences:
   "When the boat struck the rock and the water poured in and the sailors climbed up into the rigging."
3. Misspelling of common everyday words:
   "You should recieve an answer tomorrow."
4. Disagreement between subject and verb:
   "When they was told what to do, they did it."
5. Confusion as to the case of pronouns:
   "The man refused to tell either Harry or I where the money was hidden."
6. Use of unacceptable colloquial expressions:
   "He could of won the race if he had tried harder."

The purpose of Miss Barstow's test was to find out to what extent *each of her students* was likely to make *each kind of mistake* listed above.

To get an accurate picture for individual diagnosis, Miss Barstow realized that she needed a fairly long test. She wrote 20 items for each of the 6 kinds of error—a total of 120 items. Then she wrote 30 sentences that had no errors at all. She now had a total of 150 items in the test. She put all these items in random order and left enough space between them for the students to write in the necessary corrections. She told the students that, although 30 of the sentences were correct, each of the others contained a common error in either spelling, usage, or grammar. Then she allowed the students 2 fifty-minute periods to make whatever changes they thought necessary to correct the sentences.

After the students had finished the test, Miss Barstow divided the sentences into the types of error they represented and then scored each group of 20 sentences separately. Thus she had six scores for every student, one for each of the six kinds of error. The 30 correct sentences were not included in the scoring. Their function was to make it a bit more difficult for the students to identify the errors.

Using the test results, Miss Barstow was able to plan remedial work for each student's specific weaknesses.

BASIC RULES OF TEST-MAKING ILLUSTRATED BY THE FOUR TESTS

You have now read how four teachers tackled the problem of making tests. From their experiences you can see that there are basic rules which should almost always be followed:

1. Have the purpose of your test clearly in mind. To what extent are you trying to measure how well your students have learned a particular unit of study? To what extent do you hope to rank your students accurately according to their abilities? How highly diagnostic of the strengths and weaknesses of individual pupils do you want your test to be?

2. Make a careful plan for the test questions. Unless your test covers a very limited unit of work, the plan should be

written. Most plans for tests are not so simple that they can be kept firmly in mind. Furthermore, when you examine the written plan, you are better able to recognize its strengths and weaknesses.

3. If your test is mainly diagnostic in a basic skill area (as was Miss Barstow's English test), you should prepare at least 10 questions—*preferably more*—for each sub-test that you use. These sub-tests should yield separate scores on the various elements needed for mastery of the skill.

4. If you are trying to find out how well your class has mastered a particular unit of study (as was Mrs. Jackson in her fifth-grade arithmetic test), you should make a test which parallels the work in class. Generally speaking, this test should not be too difficult and the commonly accepted figure of seventy per cent for a passing score is probably appropriate for many classes.

5. When the major purpose of your test is to rank a selected group of students in order of their achievement (as was the purpose of Mr. Orlando's biology test), the questions should be on "critical" points of learning. These are the points that go beyond the superficial and obvious. They are "critical" in the sense that it is necessary to understand them for truly high-level achievement. Questions on "critical" points often require understanding implications, applying information, and reorganizing data.

For if questions are asked only on material which has been specifically taught in class and which has merely to be remembered, scores are apt to bunch near the top of the range and will not help much in determining an accurate rank order of achievement. For ranking purposes such as Mr. Orlando's, questions should go beyond what has been memorized and ask the student to use his knowledge in new situations for which it is suitable. Generally, the most accurate ranking will be achieved by making all the questions of medium difficulty for the group. But when a major purpose is to select a few of the very best students, items should be more difficult. However, do not write questions on inconsequential details just to catch students. Instead, focus on important understandings that you think the better students should have—the kinds of understandings you would not expect of poorer students. Then, even the best students are likely to miss a few questions, but their total scores will give a good estimate of their relative standings.

SPECIAL PROBLEMS IN WRITING AND SCORING TESTS

After you have decided on the purposes of the test and have made a written plan to fit these purposes, you will want to consider the following special problems about making tests and scoring them.

When Should Essay Questions Be Used and When Objective?

There are no categorical rules to tell you which type of question to use. However, it will be helpful to keep clearly in mind the characteristics of each type. Then you will be able to decide which will be most suitable for the particular purpose and circumstances of the test you are making.

The summary below compares a few of the major characteristics of the two types.

ESSAY	OBJECTIVE
Abilities Measured	
Requires the student to express himself in his own words, using information from his own background and knowledge.	Requires the student to select correct answers from given options, or to supply answer limited to one word or phrase.
Can tap high levels of reasoning such as required in inference, organization of ideas, comparison and contrast.	Can *also* tap high levels of reasoning such as required in inference, organization of ideas, comparison and contrast.
Does *not* measure purely factual information efficiently.	Measures knowledge of facts efficiently.
Scope	
Covers only a limited field of knowledge in any one test. Essay questions take so long to answer that relatively few can be answered in a given period of time. Also, the student who is especially fluent can often avoid discussing points of which he is unsure.	Covers a broad field of knowledge in one test. Since objective questions may be answered quickly, one test may contain many questions. A broad coverage helps provide reliable measurement.
Incentive to Pupils	
Encourages pupils to learn how to organize their own ideas and express them effectively.	Encourages pupils to build up a broad background of knowledge and abilities.
Ease of Preparation	
Requires writing only a few questions for a test. Tasks must be clearly defined, general enough to offer some leeway, specific enough to set limits.	Requires writing many questions for a test. Wording must avoid ambiguities and "give-aways." Distractors should embody most likely misconceptions.
Scoring	
Usually very time-consuming to score.	Can be scored quickly.
Permits teachers to comment directly on the reasoning processes of individual pupils. However, an answer may be scored differently by different teachers or by the same teacher at different times.	Answer generally scored only right or wrong, but scoring is very accurate and consistent.

How Can You Learn to Write Better Objective Questions?

A common problem is how to write objective questions that tap complex abilities. First, you need to define the kind of behavior which seems to demonstrate the ability you are trying to measure. Then ask, "How can I make up questions that will elicit this kind of behavior?"

In order to get ideas for writing your own test questions, you will find it useful to look over the questions used in published tests. You may be surprised at the ingenuity of the professional test-writer in constructing questions which measure high-level abilities.

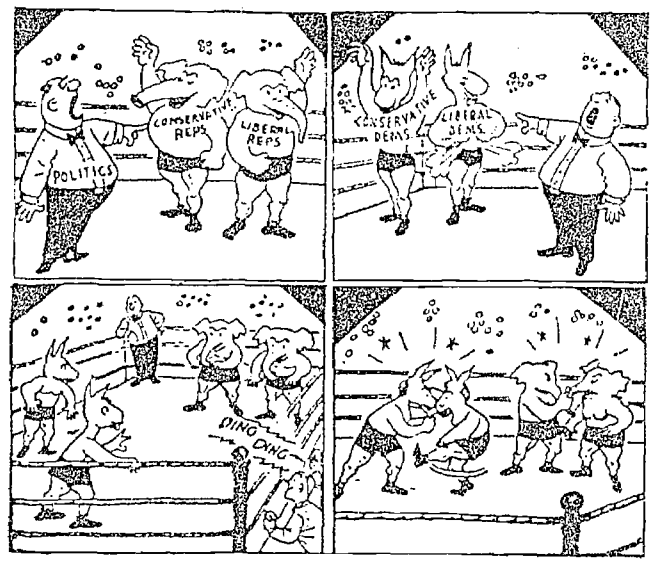
Objective questions, generally speaking, are classified into four major types: (1) Multiple-choice, (2) Matching, (3) Completion or Fill-in, (4) True-False. There are numerous variations among these types. These have been described at length in articles on measurement and in measurement textbooks (for a bibliography see page 10).

The following samples not only illustrate the four general types of objective questions, but also offer further examples of how objective questions can be constructed to require not just memorization of facts but applications of learnings and skills to new situations.

*Multiple-choice**

1. Tom wanted to find what effect fertilizer has on garden plants. He put some good soil in garden boxes. To box A he added fertilizer containing a large amount of nitrogen. To box B he added fertilizer containing a large amount of phosphorus. In each box he planted twelve bean seeds. He watered each box with the same amount of water. One thing missing from Tom's experiment was a box of soil with
   (A) both fertilizers added
   (B) neither nitrogen nor phosphorus added
   (C) several kinds of seeds planted
   (D) no seeds planted
   (Answer: B)

2. "Many young people of today are taking a great interest in the magazine Suburbia. It is of a fairly large size and of a considerable number of pages." In the second sentence above how could the size of the magazine be indicated most effectively?
   (A) By comparing it with one or two well-known magazines
   (B) By giving length, width, thickness, weight, and number of pages
   (C) By drawing a scale model
   (D) By telling how many articles each issue contains
   (Answer: A)



By permission Army Times Publishing Company

3. The cartoon illustrates which of the following characteristics of the party system in the United States?
   (A) Strong party discipline is often lacking.
   (B) The parties are responsive to the will of the voters.
   (C) The parties are often more concerned with politics than with the national welfare.
   (D) Bipartisanship often exists in name only.
   (Answer: A)

4. The situation shown in the cartoon is least likely to occur at which of the following times?
   (A) During the first session of a new Congress
   (B) During a political party convention
   (C) During a primary election campaign
   (D) During a presidential election campaign
   (Answer: D)

Matching

Read the statements below, carefully paying attention to their relation to one another. Then next to each statement mark A, B, C, or D as indicated.

(A) If the statement contains the central idea around which most of the statements can be grouped.
(B) If the statement contains a main supporting idea of the central idea.
(C) If the statement contains an illustrative fact of detailed statement related to a main supporting idea.
(D) If the statement contains an idea or ideas which are irrelevant.

1. The Roman roads connected all parts of the Empire with Rome.
2. The Roman roads were so well built that some of them remain today.
3. One of the greatest achievements of the Romans was their extensive and durable system of roads.
4. Wealthy travelers in Roman times used horse-drawn coaches.
5. Along Roman roads caravans would bring to Rome luxuries from Alexandria and the East.
6. In present-day Italy some of the roads used are original Roman roads.

(Answers: 1-B, 2-B, 3-A, 4-D, 5-C, 6-C)

*Items 1-4 by permission of the Cooperative Tests and Services, Educational Testing Service.

Completion or Fill-in

In the blank of each sentence write the word or number which best completes the sentence.

1. If people's eyes were not sensitive to blue light, objects which now appear blue would appear. . . . . .
   (Answer: black)

2. A game was played in which 28 people participated. The average final score was exactly 78. If 21 people had scores of less than 78 and 7 people had scores of more than 78 and if only whole-number scores were given, then the highest score must have been at least. . . . . .
   (Answer: 81)

True-False

In the space at the left mark whether the statement is true or false. Mark plus for true and zero for false.

. . . . . 1. If the Cascade Mountains were 500 miles farther east, western Oregon would have an increased rainfall. (Answer: +)

. . . . . 2. There is no point inside a circle farther from the edge of the circle than the length of the radius of the circle. (Answer: +)

How to avoid certain pitfalls in writing objective questions is given extensive treatment in the literature on test writing. For example, in multiple-choice or matching items, the item writer is advised to make all the choices plausible to one who does not know the correct answer. He is also advised to avoid the use of so-called "specific determiners." Examples of "specific determiners" are terms of extremes such as "always," "never," "all." A test-wise student can often guess correctly that a statement using one of these words is false. On the other hand, an extraordinarily detailed or qualified statement—to make extreme terms hold true—may also be a "specific determiner."

Two other obvious pitfalls you need to avoid in writing test questions should also be mentioned:

Reading Difficulty

Unless you are trying to test reading ability, write your questions in language that is easy for your pupils to understand.

EXAMPLE

Poor

If a man makes a business transaction wherein he purchases a motor vehicle for one thousand five hundred and twenty-five dollars and at a later date sells said vehicle in another business transaction for one thousand two hundred and sixty dollars, what is his net loss?

Better

A man buys a car for $1525 and sells it later for $1260. What is the difference between the purchase price and the selling price? (Answer: $265)

Ambiguities

Always state your questions so that there can be only one interpretation.

EXAMPLES

Poor

The shortest day of the year is in
(A) March (B) June (C) September (D) December

Better

The shortest day of the year in the northern hemisphere is in
(A) March (B) June (C) September (D) December
(Answer: D)

Poor

Which of the following books can be called humorous?
(A) *A Christmas Carol*
(B) *Tom Sawyer*
(C) *Treasure Island*
(D) *Silas Marner*
(The student might think he is expected to select more than one of the options because several of the books contain humorous incidents.)

Better

Which one of the following books is most humorous?
(A) *A Christmas Carol*
(B) *Tom Sawyer*
(C) *Treasure Island*
(D) *Silas Marner*
(Answer: B)

Some of the sources cited in the Bibliography provide numerous examples and detailed discussion of defective test items. Many of the stock admonitions appear self-evident. Yet if every item is not systematically checked, defects that later seem obvious to you may sneak by. Fortunately, you need not "go it alone" in trying to produce good questions. You will find it very helpful to work with another teacher or even a group of teachers in reviewing each other's questions. It is often surprising how easy it is to find weaknesses in questions written by someone else, and yet overlook the very same weaknesses in your own questions.

You will also find your own students are excellent critics. If you go over the questions with your students after the test, they are usually more than willing to point out ambiguities in phrasing, falseness in the keyed answer, or any other flaws.

How Can You Improve Reliability in Scoring Essay Questions?

To begin with, you should state the question in enough detail so that your pupils understand what is expected. Otherwise many of them will discuss quite different aspects of a question and their answers will vary greatly in length, points covered, and general approach. Under these conditions, you will find it difficult to compare the quality of the different answers and assign grades accurately.

Below is an example of an essay question which is too general, followed by an example of one which is stated in greater detail:

EXAMPLE

Poor

Explain why you think the United Nations has been a success or a failure.

Better

An important function of the United Nations is to help settle disputes between nations. Describe how one dispute was handled successfully, pointing out how the settlement illustrates a general strength of the United Nations. Describe also how one dispute was handled unsuccessfully, pointing out how this illustrates a general weakness of the United Nations. Your essay should be about 300-400 words in length (2 or 3 pages in longhand).

Various systematic procedures have been set up to make the scoring of essay questions more reliable. These procedures are useful but, unfortunately, time-consuming. You will have to decide how conscientiously you wish to follow these procedures—whether the increase in reliability is worth the additional time and effort.

The following method for scoring essay questions is described by Graham.* It is clear cut and relatively simple. You may find it useful.

(1) The teacher *analyzes* the points that he thinks should be made in the ideal response *and assigns a numerical weight* to each point. Some points may be of greater importance; hence, they would be weighted appropriately. The instructor may wish to allow extra credit for clear organization of thinking. Sometimes he may feel that he cannot develop a "scoring key" until he reads a cross-section of students' papers. Whether derived by teacher-analysis, or by analysis of pupil responses, or by a combination of the two approaches, a systematic method of scoring using numerical values or percentages increases objectivity.

(2) The test reader *evaluates all the responses to one question* before going on to score the next question.

(3) As the teacher reads, he *tosses the papers into five piles* (high to low in quality). This procedure may be unnecessary if the instructor is satisfied with the quantitative appraisal described in (1) above; but if he also wants a qualitative estimate, he may need to re-check his classifications to determine if the papers in each pile are indeed of similar quality.

(4) *Anonymity is necessary* for the accurate scoring of essay tests because of the ubiquitous "halo effect." The easiest way to prevent this kind of subjectivity is to ask pupils to write their names *only* on the back of their test papers.

Since it is difficult to grade essays reliably, you will usually be more concerned with writing comments than awarding grades. Your written comment on a paper will help the student more than a grade in understanding his strengths and weaknesses. However, if essay tests must be used for grades, reliability can be increased by basing the final grade on several essay tests rather than one.

*See Bibliography (6)

What Kind of Statistical Analysis of Test Questions Should the Teacher Make?

The very words "statistics" causes some teachers to shudder and feel lost. This need not be. Practical statistics can be quite simple. Of course few teachers are willing to spend time on statistical analysis of a question that is not to be used again. However, many teachers are well aware of the advantages of a copious stockpile of questions. If certain characteristics of these questions are known, it is possible to design a test more effectively for a specific purpose and group.

Two simplified procedures by which teachers can obtain such information about their test items have been described in detail in two publications listed in the Bibliography. Diederich (4) urges pupil participation in the analysis of test items, emphasizing instructional values. Katz (7) shows how the classroom teacher can complete an item analysis in remarkably little time. In general, item analysis tells you two things that you will want to know about the questions you are keeping in your stockpile: (1) how difficult each question is; (2) how well each question discriminates between high- and low-ranking students on the test as a whole.

Difficulty

A simple measure of difficulty is the per cent of students who got the question right on any one test. For example, the first time you use a particular question, 75 per cent of the students may get it right. If you keep a card file of your questions, then on the back of the card for this question you write the date and "75% right." You make a similar entry every time you use the question. Then, knowing in advance the probable difficulty of the questions, you will be able to "tailor" your test (in a style determined by your purpose) to fit the ability level of your class.

As the basic rules (pages 4–5) pointed out, when you want to see whether your class has mastered a fundamental unit of study, questions will tend to be easy. But when the principal purpose of the test is to rank all the students in the group in order of ability—for example, to give them grades which reflect standing in class—try to use questions which are of middle difficulty.* If you want to discriminate more precisely among the better students (as Mr. Orlando did), all of the questions should be of greater difficulty. In fact, test-makers speak of "peaking a test at the cut-off point"—that is, trying to make each test item so difficult that (on a free-response test) only about 50% of the lowest pupils selected or highest pupils rejected will answer it correctly.

Of course, in writing difficult questions, you will take care that the difficulty is based on significant parts of the class-

*Ideally, for this purpose, each question on a free-response test should be so difficult that only 50% of the group get it right; on a 5-option multiple choice test, 60%; 4-option multiple-choice, 62%; 3-option, 66%; true-false, 75%—assuming that the score is the number of items answered correctly.

work, which your better students should be able to understand. As has already been emphasized, the difficulty should not be based on memorization of trivial details or result from tricky or ambiguous wording. In other words, your questions should be fair and truly representative of important teaching objectives. Very often questions that require pupils to apply information learned in the classroom to new situations are good for this purpose. The science test that Mr. Orlando made for his biology class included many questions of this type.

In connection with difficulty, most tests should give your students sufficient time to consider each question and answer it to the best of their abilities. Probably about 90 per cent of the students should complete all the questions on a test. Of course, if your purpose is to test speed itself, then you will expect fewer students to finish the test. For example, if you were testing speed of reading, you might expect only 10 to 15 per cent to answer all the questions.

Discrimination

You will want to know how effectively each question contributes to the discrimination between the high-scoring and low-scoring pupils on the test as a whole. If an important purpose of your test is to rank students according to ability, some questions will help more than others. The discriminating power of each question can be estimated by these steps:

1. Arrange the test papers in order of scores, with the highest score on top.
2. Take a specified quantity (ten are easy to handle) of the papers from the top and the same quantity from the bottom. Place them in separate piles called High and Low, respectively.
3. Now you are ready to analyze the individual questions. For each question count the number of Highs who got it right and the number of Lows who got it right.
4. Convert these numbers to per cents. If the question is a good one for ranking students, then substantially more of the Highs than Lows will have answered it correctly.

For example, assume there are 10 students in the top group and 10 in the bottom group. On the first question, 8 of the Highs (80%) but only 3 of the Lows (30%) got the right answer. This is a good question for ranking students because it is clear that the students who generally did well on the test were able to get the question right, while those who did poorly on the test as a whole tended to get it wrong. The per cent of the top group and the per cent of the bottom group getting the item right may be written on the back of the question card in your stockpile of items. In the example given above you would note H=80%, L=30%. Later you may refer to this information in assembling a test which will be especially effective in discrimination.

You will want to take a close and suspicious "second look" at any question where the top students had as much difficulty as the poorer students—or more. Possibly the question is not so clearly stated as it should be. If it is a multiple-choice item, perhaps one of the "wrong" options is too close to being correct. Look over each question where the results of the analysis indicate the possibility of a flaw in the question and decide whether you can improve the question by rewriting.

When Should You Use a Published Test and When Should You Make Your Own?

If a published achievement test *covers the points* you wish to measure, you may use it instead of making your own test. Sometimes, however, it is impossible to find a published test that matches what you want to measure—particularly if you want a test covering a single unit of study. Thus, when you need a test covering only a few weeks of class work, you will ordinarily make your own. One important benefit from preparing your own test is that the very process of writing questions forces you to define your own teaching objectives in terms of specific skills and understandings.

However, there are many times when you may decide to use a published test. And, of course, your school probably gives standardized tests as part of its testing program. These tests offer several advantages. They have been written by specialists so that the general quality of questions is high. They have been subjected to careful statistical analysis so that the questions are controlled for difficulty and discrimination. They are accompanied by norms so that you are able to compare the performance of your pupils with the performance of a representative sample with known characteristics.

And finally, the fact that well-constructed achievement tests have been prepared by groups of experienced teachers, whose competence is generally recognized, gives you a check on your judgment of what should be measured. Their consensus on skills and understandings to be covered is not necessarily better than yours. But if it is substantially different from yours, it provides a valuable supplement to your estimates of student achievement in any area. It gives some notion of how well your pupils can do on a test of learnings that teachers in general may consider important.

A Word in Closing

The most scrupulous heed for all the cautions, admonitions, principles, and procedures discussed in this pamphlet will not guarantee that you will make good tests—although it may do much to prevent bad ones. Good tests cannot be written merely by following any set of "rules." There is an art to good test-writing which involves elements of originality and creativity as well as knowledge of theory. This pamphlet does not pretend to provide competence in all elements of the art. Such competence seems to thrive on practice, criticism, tryout, analysis, and more practice.

However, this pamphlet does attempt to foster an attitude and approach which have helped many teachers to improve their tests. It has presented some essential principles and some realistic illustrations to serve as guides and touchstones in your efforts to make better classroom tests.

Bibliography

1. Anderson, Scarvia B., Katz, M. R., and Shimberg, B. *Meeting the Test* (Revised Edition). New York: The Four Winds Press, 1965.
2. Baron, D. and Bernard, H. W. *Evaluation Techniques for Classroom Teachers*. New York: McGraw-Hill, 1958. Chapter 12.
3. Bean, K. L. *Construction of Educational and Personal Tests*. New York: McGraw-Hill, 1953.
4. Diederich, P. *Short-cut Statistics for Teacher-made Tests*. Princeton, N.J.: Educational Testing Service, 1960.
5. Furst, E. J. *Constructing Evaluation Instruments*. New York: Longmans, Green and Co., 1958.
6. Graham, Grace. "Teachers Can Construct Better Achievement Tests." *Curriculum Bulletin*. University of Oregon, Vol. XII, No. 170, December, 1956.
7. Katz, M. "Improving Classroom Tests by Means of Item Analysis." *The Clearing House*. January, 1961, 265-269.
8. Odell, C. W. *How To Improve Classroom Testing*. Dubuque, Iowa: William C. Brown, 1953.
9. Thomas, R. M. *Judging Student Progress* (2nd Edition). New York: Longmans, Green and Co., 1960. Chapter 3.
10. Thorndike, R. L. and Hagen, Elizabeth. *Measurement and Evaluation in Psychology and Education* (2nd Edition). New York: John C. Wiley and Sons, 1961. Chapters 3 and 4.
11. Wesman, A. G. "Writing the Test Item." In R. L. Thorndike (Ed.), *Educational Measurement* (2nd Edition). Washington, D.C.: American Council on Education, 1971. Chapter 4.
12. Wood, Dorothy Adkins. *Test Construction: Development and Interpretation of Achievement Tests*. Columbus, Ohio: C. E. Merrill, 1960.