How to Prepare Better

Multiple-Choice Test Items:

Guidelines for University Faculty

Steven J. Burton
Richard R. Sudweeks
Paul F. Merrill
Bud Wood

Brigham Young University Testing Services
and
The Department of Instructional Science

TABLE OF CONTENTS

Introduction

Have you ever seen a truly awful multiple-choice test question? One that is so defective that the correct answer is either obvious, debatable, obscure, or missing altogether? One that makes you wonder what the test writer had in mind when he or she constructed it? The following is such a question:

> Technicle advances in farm equipment; a. encourage urbanization because fewer people live on farms b. higher food prices c. revolutionizd the industry d never occurs rapidly e. both a and c d. none of the above

Most multiple-choice test questions are not as replete with errors as this example, but you have probably seen many of the errors before. In addition to confusing and frustrating students, poorly-written test questions yield scores of dubious value that are inappropriate to use as a basis of evaluating student achievement. Compare the example above with the following one:

> Which of the following is the best explanation of why technical advances in farm equipment led to an increase in urbanization?
>
> a. Fewer people were needed to run the farms.
> b. Fewer people were qualified to operate the equipment.
> c. More people could live in the city and commute to the farm.
> d. More people went to work at the equipment manufacturing plants.

While this example may still leave room for improvement, it is certainly superior to the first one. Well-written multiple-choice test questions do not confuse students, and yield scores that are more appropriate to use in determining the extent to which students have achieved educational objectives.

Booklet Objectives

Most poorly-written multiple-choice test questions are characterized by at least one of the following three weaknesses:

- They attempt to measure an objective for which they are not well-suited
- They contain clues to the correct answer
- They are worded ambiguously

Well-written test questions (hereafter referred to as test *items*) are defined as those that are constructed in adherence to guidelines designed to avoid the three problems listed above.

The purpose of this booklet is to present those guidelines with the intent of improving the quality of the multiple-choice test items used to assess student achievement. Specifically, the booklet is designed to help teachers achieve the following objectives:

1. Distinguish between objectives which can be appropriately assessed by using multiple-choice items and objectives which would be better assessed by some other means.

2. Evaluate existing multiple-choice items by using commonly-accepted criteria to identify specific flaws in the items.

3. Improve poorly-written multiple-choice items by correcting the flaws they contain.

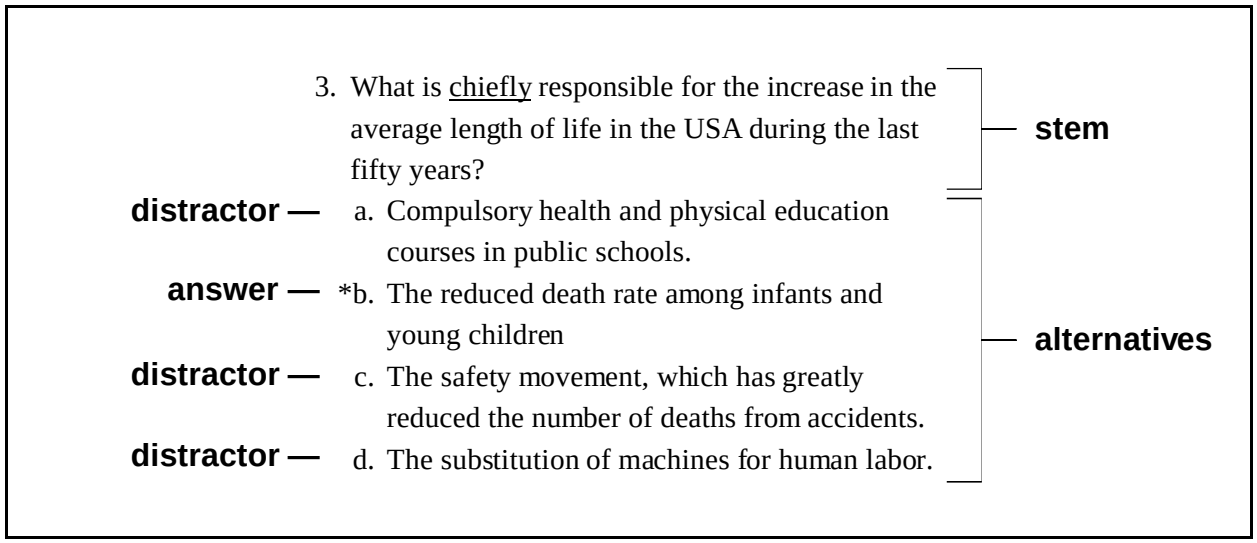
4. Construct well-written multiple-choice items that measure given objectives.

Anatomy of a Multiple-Choice Item

A standard multiple-choice test item consists of two basic parts: a problem (*stem*) and a list of suggested solutions (*alternatives*). The stem may be in the form of either a question or an incomplete statement, and the list of alternatives contains one correct or best alternative (*answer*) and a number of incorrect or inferior alternatives (*distractors*).

The purpose of the distractors is to appear as plausible solutions to the problem for those students who have not achieved the objective being measured by the test item. Conversely, the distractors must appear as *im*plausible solutions for those students who *have* achieved the objective. Only the answer should appear plausible to these students.

In this booklet, an asterisk (*) is used to indicate the answer.



Advantages and Limitations of Multiple-Choice Items

Multiple-choice test items are not a panacea. They have advantages and limitations just as any other type of test item. Teachers need to be aware of these characteristics in order to use multiple-choice items effectively.

Advantages

Versatility. Multiple-choice test items are appropriate for use in many different subject-matter areas, and can be used to measure a great variety of educational objectives. They are adaptable to various levels of learning outcomes, from simple recall of knowledge to more complex levels, such as the student's ability to:

- Analyze phenomena
- Apply principles to new situations
- Comprehend concepts and principles
- Discriminate between fact and opinion
- Interpret cause-and-effect relationships
- Interpret charts and graphs
- Judge the relevance of information
- Make inferences from given data
- Solve problems

The difficulty of multiple-choice items can be controlled by changing the alternatives, since the more homogeneous the alternatives, the finer the distinction the students must make in order to identify the correct answer. Multiple-choice items are amenable to item analysis, which enables the teacher to improve the item by replacing distractors that are not functioning properly. In addition, the distractors chosen by the student may be used to diagnose misconceptions of the student or weaknesses in the teacher's instruction.

Validity. In general, it takes much longer to respond to an essay test question than it does to respond to a multiple-choice test item, since the composing and recording of an essay answer is such a slow process. A student is therefore able to answer many multiple-choice items in the time it would take to answer a single essay question. This feature enables the teacher using multiple-choice items to test a broader sample of course content in a given amount of testing

time. Consequently, the test scores will likely be more representative of the students' overall achievement in the course.

Reliability. Well-written multiple-choice test items compare favorably with other test item types on the issue of reliability. They are less susceptible to guessing than are true-false test items, and therefore capable of producing more reliable scores. Their scoring is more clear-cut than short-answer test item scoring because there are no misspelled or partial answers to deal with. Since multiple-choice items are objectively scored, they are not affected by scorer inconsistencies as are essay questions, and they are essentially immune to the influence of bluffing and writing ability factors, both of which can lower the reliability of essay test scores.

Efficiency. Multiple-choice items are amenable to rapid scoring, which is often done by scoring machines. This expedites the reporting of test results to the student so that any follow-up clarification of instruction may be done before the course has proceeded much further. Essay questions, on the other hand, must be graded manually, one at a time.

Limitations

Versatility. Since the student selects a response from a list of alternatives rather than supplying or constructing a response, multiple-choice test items are not adaptable to measuring certain learning outcomes, such as the student's ability to:

- Articulate explanations
- Display thought processes
- Furnish information
- Organize personal thoughts
- Perform a specific task
- Produce original ideas
- Provide examples

Such learning outcomes are better measured by short answer or essay questions, or by performance tests.

Reliability. Although they are less susceptible to guessing than are true false-test items, multiple-choice items are still affected to a certain extent. This guessing factor reduces the

reliability of multiple-choice item scores somewhat, but increasing the number of items on the test offsets this reduction in reliability. The following table illustrates this principle.

Number of 4-Alternative Multiple-Choice Items on Test	Chance of Scoring 70% or Higher by Blind Guessing Alone
2	1 out of 16
5	1 out of 64
10	1 out of 285
15	1 out of 8,670
20	1 out of 33,885
25	1 out of 942,651

For example, if your test includes a section with only two multiple-choice items of 4 alternatives each (*a b c d*), you can expect 1 out of 16 of your students to correctly answer both items by guessing blindly. On the other hand if a section has 15 multiple-choice items of 4 alternatives each, you can expect only 1 out of 8,670 of your students to score 70% or more on that section by guessing blindly.

Difficulty of Construction. Good multiple-choice test items are generally more difficult and time-consuming to write than other types of test items. Coming up with plausible distractors requires a certain amount of skill. This skill, however, may be increased through study, practice, and experience.

Deciding When Multiple-Choice Items Should Be Used

In order for scores to accurately represent the degree to which a student has attained an educational objective, it is essential that the form of test item used in the assessment be suitable for the objective. Multiple-choice test items are often advantageous to use, but they are not the best form of test item for every circumstance. *In general, they are appropriate to use when the attainment of the educational objective can be measured by having the student select his or her response from a list of several alternative responses.*

- If the attainment of the educational objective can be better measured by having the student *supply* his response, a **short-answer item** or **essay question** may be appropriate.

- If there are several homogeneous test items, it may be possible to combine them into a single **matching item** for more efficient use of testing time.

- If the attainment of the objective can be better measured by having the student *do* something, a **performance test** should be considered.

Examples

The following table contains sample educational objectives in the first column. The second column indicates whether or not multiple-choice items are appropriate for measuring the attainment of the objectives in the first column. The third column states the reason for the response in the second column, and a suggested item type to use if the response is *no*.

Educational Objective	Multiple-Choice Item Okay?	Reason (Appropriate item type if not Multiple Choice)
Writes complete sentences	No	Response must be supplied. (Essay)
Identifies errors in punctuation	Yes	Response may be selected.
Expresses own ideas clearly	No	Response must be supplied. (Essay)
Uses gestures appropriately in giving a speech	No	Response must be supplied. (Performance)
Identifies the parts of a sentence	Yes	Response may be selected.

Measuring Higher-Level Objectives with Multiple-Choice Items

One of the reasons why some teachers dislike multiple-choice items is that they believe these items are only good for measuring simple recall of facts. This misconception is understandable, because multiple-choice items are frequently used to measure lower-level objectives, such as those based on knowledge of terms, facts, methods, and principles. The real value of multiple-choice items, however, is their applicability in measuring higher-level objectives, such as those based in comprehension, application, and analysis.

Examples

Comprehension

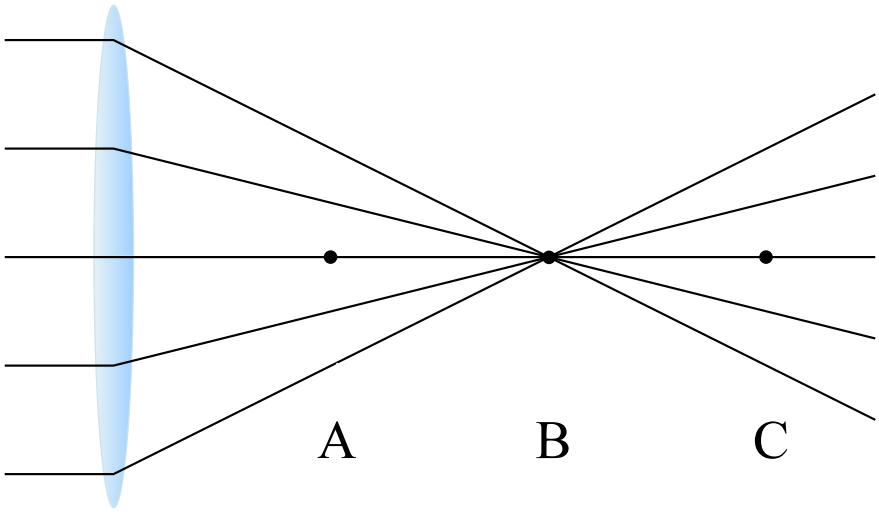
Objective: Identifies the effect of changing a parameter (rule using).

A pendulum consists of a sphere hanging from a string. What will happen to the period of the pendulum if the mass of the sphere is doubled? (Assume that the effects of air friction and the mass of the string are negligible, and that the sphere traces an arc of 20° in a plane as it swings.)

a. It will increase.
b. It will decrease.
*c. It will remain unchanged.
d. More information is needed to determine what will happen.

Application

Objective: Identifies the correct application of principle (problem solving).



In the diagram above, parallel light rays pass through a convex lens and converge to a focus. They can be made parallel again by placing a:

a. Concave lens at point B.
b. Concave lens at point C.
c. Second convex lens at point A.
d. Second convex lens at point B.
*e. Second convex lens at point C.

Analysis

Objective: Analyzes poetry and identifies patterns and relationships.

[The poem is included here.]

The chief purpose of stanza 9 is to:

a. Delay the ending to make the poem symmetrical.
b. Give the reader a realistic picture of the return of the cavalry.
c. Provide material for extending the simile of the bridge to a final point.
*d. Return the reader to the scene established in stanza 1.

Varieties of Multiple-Choice Items

Single Correct Answer

In items of the *single-correct-answer* variety, all but one of the alternatives are incorrect; the remaining alternative is the correct answer. The student is directed to identify the correct answer.

Example

A market clearing price is a price at which:
a. Demand exceeds supply.
*b. Supply equals demand.
c. Supply exceeds demand.

Best Answer

In items of the *best-answer* variety, the alternatives differ in their degree of correctness. Some may be completely incorrect and some correct, but one is clearly more correct than the others. This best alternative serves as the answer, while the other alternatives function as distractors. The student is directed to identify the best answer.

Example

Monopolies cause problems in a market system because they:
a. Create external costs and imperfect information.
*b. Lead to higher prices and under production.
c. Make such large profits.
d. Manufacture products of poor quality.

Negative

In items of the *negative* variety, the student is directed to identify either the alternative that is an incorrect answer, or the alternative that is the *worst* answer. Any of the other multiple-choice varieties can be converted into this negative format.

Example

Which of the following is *not* true of George Washington?
a. He served only two terms as president.
b. He was an experienced military officer before the Revolutionary War.
c. He was born in 1732.
*d. He was one of the signers of the Declaration of Independence.

For most educational objectives, a student's achievement is more effectively measured by having him or her identify a correct answer rather than an incorrect answer. Just because the student knows an incorrect answer does not necessarily imply that he or she knows the correct answer. For this reason, *items of the negative variety are not recommended for general use.*

Occasionally, negative items are appropriate for objectives dealing with health or safety issues, where knowing what not to do is important. In these situations, negative items must be carefully worded to avoid confusing the student. The negative word should be placed in the stem, not in the alternatives, and should be emphasized by using underlining, *italics*, **bold face**, or CAPITALS. In addition, each of the alternatives should be phrased positively to avoid forming a confusing double negative with the stem.

Poor Example

All of the following are correct procedures for putting out a fire in a pan on the stove except:
a. Do not move the pan.
*b. Pour water into the pan.
c. Slide a fitted lid onto the pan.
d. Turn off the burner controls.

Better Example

All of the following are correct procedures for putting out a fire in a pan on the stove *except*:
a. Leave the pan where it is.
*b. Pour water into the pan.
c. Slide a fitted lid onto the pan.
d. Turn off the burner controls.

The negative word "except" in the poor example above is not emphasized, and alternative *a* forms a double negative with the stem. These defects have been corrected in the better example.

Research. In a survey of 46 authoritative references in the field of educational measurement, 31 of the 35 authors that discussed the negative variety recommend that they be avoided (Haladyna & Downing, 1989a).

Multiple Response

In items of multiple response variety, two or more of the alternatives are keyed as correct answers; the remaining alternatives serve as distractors. The student is directed to identify each correct answer.

Example

Which of the following is a characteristic of a virus?
*a. It can cause disease.
b. It can reproduce by itself.
c. It is composed of large living cells.
*d. It lives in plant and animal cells.

This variety of item can be scored in several different ways. Scoring on an all-or-none basis (one point if all the correct answers and none of the distractors are selected, and zero points otherwise), and scoring each alternative independently (one point for each correct answer chosen and one point for each distractor not chosen) are commonly used methods. Both methods, however, have distinct disadvantages. With the first method, a student who correctly identifies all but one of the answers receives the same score as a student who cannot identify any of the answers.

The second method produces scores more representative of each student's achievement, but most computer programs currently used with scoring machines do not include this method as an option. As a result, *items of the multiple-response variety are not recommended.*

Since an item of multiple-response variety is often simply a series of related true-false questions presented together as a group, a good alternative that avoids the scoring problems mentioned above is to rewrite it as a multiple true-false item.

Multiple True-False Example

A virus:

*T F	Can cause disease.
T *F	Can reproduce by itself.
T *F	Is composed of large living cells.
*T F	Lives in plant and animal cells.

Research. In a survey of 46 authoritative references in the field of educational measurement, 32 of the 35 authors that discussed how many correct answers to include in an item recommend using only one (Haladyna & Downing, 1989a). In addition, items of the multiple-response variety have been found to be lower in reliability, higher in difficulty, and equal in validity when compared with similar multiple true-false items (Frisbie, 1990).

Combined Response

In items of the combined-response variety, one or more of the alternatives are correct answers; the remaining alternatives serve as distractors. The student is directed to identify the correct answer or answers by selecting one of a set of letters, each of which represent a combination of alternatives.

Example

The fluid imbalance known as edema is commonly associated with:

1. Allergic reactions.
2. Congestive heart failure.
3. Extensive burns.
4. Protein deficiency.

The correct answer is:

a. 1, 2, and 3.
b. 1 and 3.
c. 2 and 4.
d. 4 only.
*e. 1, 2, 3, and 4.

This variety is also known as complex multiple-choice, multiple multiple-choice, or type K. It shares the disadvantage of all-or-none scoring with the multiple-response variety discussed previously, and has the added disadvantage of providing clues that help students with only partial knowledge detect the correct combination of alternatives. In the example above, a student can

identify combination *e* as the correct response simply by knowing that alternatives 1 and 4 are both correct. Because of these disadvantages, *items of combined-response variety are not recommended.*

Like the multiple-response variety, an item of the combined-response variety is often simply a series of related true-false questions presented together as a group. A good alternative that avoids the scoring and cluing problems mentioned above is to rewrite it as a multiple true-false item.

Multiple True-False Example

The fluid imbalance known as edema is commonly associated with:

*T F allergic reactions.
*T F congestive heart failure.
*T F extensive burns.
*T F protein deficiency.

Research. Numerous studies indicate that items of the combined-response variety are lower in reliability, lower in discrimination, higher in difficulty, and equal in validity when compared with similar items of the single-correct-answer and best-answer varieties (Albanese, 1990; Haladyna & Downing, 1989b). They have also been found to be lower in reliability, higher in difficulty, and equal in validity when compared with similar multiple true-false items (Frisbie, 1990).

Guidelines for Constructing Multiple-Choice Items

The following guidelines and the checklist found inside the back cover of this booklet will help you construct better multiple-choice test items. These guidelines are specifically designed for the single-answer and best-answer varieties of multiple-choice items.

1. Construct each item to assess a single written objective.

Items that are not written with a specific objective in mind often end up measuring lower-level objectives exclusively, or covering trivial material that is of little educational worth.

Research. Although few studies have addressed this issue, one study has found that basing items on objectives makes the items easier and more homogeneous (Baker, 1971).

2. Base each item on a specific problem stated clearly in the stem.

The stem is the foundation of the item. After reading the stem, the student should know exactly what the problem is and what he or she is expected to do to solve it. If the student has to infer what the problem is, the item will likely measure the student's ability to draw inferences from vague descriptions rather than his or her achievement of a course objective.

Poor Example

California:

a. Contains the tallest mountain in the United States
b. Has an eagle on its state flag.
c. Is the second largest state in terms of area.
*d. Was the location of the Gold Rush of 1849.

Better Example

What is the main reason so many people moved to California in 1849?

a. California land was fertile, plentiful, and inexpensive.
*b. Gold was discovered in central California
c. The east was preparing for a civil war.
d. They wanted to establish religious settlements.

As illustrated in the following examples, the stem may consist of either a direct question or an incomplete sentence, whichever presents the problem more clearly and concisely.

Direct Question Example

Which of the following was the principal keyboard instrument in 16th century Europe?
a. Clavichord.
*b. Harpsichord.
c. Organ.
d. Pianoforte.

Incomplete Sentence Example

The principal keyboard instrument in 16th century Europe was the:
a. Clavichord.
*b. Harpsichord.
c. Organ.
d. Pianoforte.

3. Include as much of the item as possible in the stem, but do not include irrelevant material.

Rather than repeating redundant words or phrases in each of the alternatives, place such material in the stem to decrease the reading burden and more clearly define the problem in the stem.

Poor Example

If the pressure of a certain amount of gas is held constant, what will happen if its volume is increased?
a. The temperature of the gas will decrease.
*b. The temperature of the gas will increase.
c. The temperature of the gas will remain the same.

Better Example

If you increase the volume of a certain amount of gas while holding its pressure constant, its temperature will:
a. Decrease.
*b. Increase.
c. Remain the same.

Notice how the underlined words are repeated in each of the alternatives in the poor example above. This problem is fixed in the better example, where the stem has been reworded to include the words common to all of the alternatives.

Excess material in the stem that is not essential to answering the problem increases the reading burden and adds to student confusion over what he or she is being asked to do.

Poor Example

Suppose you are a mathematics professor who wants to determine whether or not your teaching of the unit on probability has had a significant effect on your students. You decide to analyze their scores from a test they took before the instruction and their scores from another exam taken after the instruction. Which of the following t-tests is appropriate to use in this situation?

*a. Dependent samples.
b. Heterogeneous samples.
c. Homogeneous samples.
d. Independent samples.

Better Example

When analyzing your students' pretest and posttest scores to determine if your teaching has had a significant effect, an appropriate statistic to use is the t-test for:

*a. Dependent samples.
b. Heterogeneous samples.
c. Homogeneous samples.
d. Independent samples.

The stem of the poor example above is excessively long for the problem it is presenting. The stem of the better example has been reworded to exclude most of the irrelevant material, and is less than half as long.

Research. Several studies have indicated that including irrelevant material in the item stem decreases both the reliability and the validity of the resulting test scores (Haladyna & Downing, 1989b).

4. State the stem in positive form (in general).

Negatively-worded items are those in which the student is instructed to identify the exception, the incorrect answer, or the least correct answer. Such items are frequently used, because they are relatively easy to construct. The teacher writing the item need only come up with one distractor, rather than the two to four required for a positively-worded item.

Positive items, however, are more appropriate to use for measuring the attainment of most educational objectives. For information on appropriate uses of negative items, see the section in this booklet entitled, "Varieties of Multiple-Choice Items."

5. Word the alternatives clearly and concisely.

Clear wording reduces student confusion, and concise wording reduces the reading burden placed on the student.

Poor Example

The term *hypothesis*, as used in research, as defined as:

a. A conception or proposition formed by speculation or deduction or by abstraction and generalization from facts, explaining or relating an observed set of facts, given probability by experimental evidence or by factual or conceptual analysis but not conclusively established or accepted.
b. A statement of an order or relation of phenomena that so far as is known is invariable under the given conditions, formulated on the basis of conclusive evidence or tests and universally accepted, that has been tested and proven to conform to facts.
*c. A proposition tentatively assumed in order to draw out its logical or empirical consequences and so test its accord with facts that are known or may be determined, of such a nature as to be either proved or disproved by comparison with observed facts.

Better Example

The term hypothesis, as used in research, is defined as:

a. An assertion explaining an observed set of facts that has not been conclusively established.
b. A universally accepted assertion explaining an observed set of facts.
*c. A tentative assertion that is either proved or disproved by comparison with an observed set of facts.

The alternatives in the poor example above are rather wordy, and may require more than one reading before the student understands them clearly. In the better example, the alternatives have been streamlined to increase clarity without losing accuracy.

6. Keep the alternatives mutually exclusive.

Alternatives that overlap create undesirable situations. Some of the overlapping alternatives may be easily identified as distractors. On the other hand, if the overlap includes the intended answer, there may be more than one alternative that can be successfully defended as being the answer.

Poor Example

How long does an *annual* plant generally live?
- *a. It dies after the first year.
- b. It lives for many years.
- c. It lives for more than one year.
- *d. It needs to be replanted each year.

Better Example

How long does an *annual* plant generally live?
- *a. Only one year.
- b. Only two years.
- c. Several years.

In the poor example above, alternatives *a* and *d* overlap, as do alternatives *b* and *c*. In the better example, the alternatives have been rewritten to be mutually exclusive.

7. Keep the alternatives homogeneous in content.

If the alternatives consist of a potpourri of statements related to the stem but unrelated to each other, the student's task becomes unnecessarily confusing. Alternatives that are parallel in content help the item present a clear-cut problem more capable of measuring the attainment of a specific objective.

Poor Example

Idaho is widely known as:
- *a. The largest producer of potatoes in the United States.
- b. The location of the tallest mountain in the United States.
- c. The state with a beaver on its flag.
- d. The "Treasure State."

Better Example

Idaho is widely known for its:
- a. Apples.
- b. Corn.
- *c. Potatoes.
- d. Wheat.

The poor example contains alternatives testing knowledge of state agriculture, physical features, flags, and nicknames. If the student misses the item, it does not tell the teacher in which of the four areas the student is weak. In the better example, all of the alternatives refer to state agriculture, so if the student misses the item, it tells the teacher that the student has a weakness in that area.

8. Keep the alternatives free from clues as to which response is correct.

Poorly-written items often contain clues that help students who do not know the correct answer eliminate incorrect alternatives and increase their chance of guessing correctly. Such items tend to measure how clever the students are at finding the clues rather than how well they have attained the objective being measured. The following suggestions will help you detect and remove many of these clues from your items.

8.1 Keep the grammar of each alternative consistent with the stem. Students often assume that inconsistent grammar is the sign of a distractor, and they are generally right.

Poor Example

A word used to describe a noun is called an:
*a. Adjective.
b. Conjunction.
c. Pronoun.
d. Verb.

Better Example

A word used to describe a noun is called:
*a. An adjective.
b. A conjunction.
c. A pronoun.
d. A verb.

The word "an" in the stem of the poor example above serves as a clue to the correct answer, "adjective," because the other alternatives begin with consonants. The problem has been corrected in the better example by placing the appropriate article, "an" or "a," in each alternative.

Poor Example

Which of the following would do the most to promote the application of nuclear discoveries to medicine?
a. Trained radioactive therapy specialists.
*b. Developing standardized techniques for treatment of patients.
c. Do not place restrictions on the use of radioactive substances.
d. If the average doctor is trained to apply radioactive treatments.

Better Example

Which of the following would do the most to promote the application of nuclear discoveries to medicine?
a. Adding trained radioactive therapy specialists to hospital staffs.
*b. Developing standardized techniques for treatment patients.
c. Removing restrictions on the use of radioactive substances.
d. Training the average doctor to apply radioactive treatments.

In the poor example above, the answer fits better grammatically with the stem than do the distractors. This problem has been solved in the better example by rewording the alternatives.

Research. Several studies have found that grammatical clues make items easier (Haladyna & Downing, 1989b).

8.2 Keep the alternatives parallel in form. If the answer is worded in a certain way and the distractors are worded differently, the student may take notice and respond accordingly.

Poor Example

You have just spent ten minutes trying to teach one of your new employees how to change a typewriter ribbon. The employee is still having a great deal of difficulty changing the ribbon, even though you have always found it simple to do. At this point, you should:

a. Tell the employee to ask an experienced employee working nearby to change the ribbon in the future.
b. Tell the employee that you never found this difficult, and ask what he or she finds difficult about it.
*c. Review each of the steps you have already explained, and determine whether the employee understands them.
d. Tell the employee that you will continue teaching him or her later, because you are becoming irritable.

Better Example

You have just spent ten minutes trying to teach one of your new employees how to change a typewriter ribbon. The employee is still having a great deal of difficulty changing the ribbon, even though you have always found it simple to do. At this point, you should:

a. Ask an experienced employee working nearby to change the ribbon in the future.
b. Mention that you never found this difficult, and ask what he or she finds difficult about it.
*c. Review each of the steps you have already explained, and determine whether the employee understands them.
d. Tell the employee that you will continue teaching him or her later because you are becoming irritable.

The answer in the poor example above stands out because it does not include the identical wording underlined in each of the distractors. The answer is less obvious in the better example because the distractors have been reworded to be more parallel with the answer.

8.3 Keep the alternatives similar in length. An alternative noticeably longer or shorter than the other is frequently assumed to be the answer, and not without good reason.

Poor Example

Which of the following is the best indication of high morale in a supervisor's unit?
a. The employees are rarely required to work overtime.
*b. The employees are willing to give first priority to attaining group objectives, subordinating any personal desires they may have.
c. The supervisor enjoys staying late to plan the next day.
d. The unit gives expensive birthday presents to each other.

Better Example

Which of the following is the best indication of high morale in a supervisor's unit?
a. The employees are rarely required to work overtime.
*b. The employees willingly give first priority to attaining group objectives.
c. The supervisor enjoys staying late to plan for the next day.
d. The unit members give expensive birthday presents to each other.

Notice how the answer stands out in the poor example above. Both the answer and one of the distractors have been reworded in the better example to make the alternative lengths more uniform.

Research. Numerous studies have indicated that items are easier when the answer is noticeably longer than the distractors when all of the alternatives are similar in length (Haladyna & downing, 1989b).

8.4 Avoid textbook, verbatim phrasing. If the answer has been lifted word-for-word from the pages of the textbook, the students may recognize the phrasing and choose correctly out of familiarity rather than achievement.

Poor Example

The term *operant conditioning* refers to the learning situation in which:
a. A familiar response is associated with a new stimulus.
b. Individual associations are linked together in sequence.
*c. A response of the learner is instrumental in leading to a subsequent reinforcing event.
d. Verbal responses are made to verbal stimuli.

Better Example

The term *operant conditioning* refers to the learning situations in which:
a. A familiar response is associated with a new stimulus.
b. Individual associations are linked together in sequence.
*c. The learner's response leads to reinforcement.
d. Verbal responses are made to verbal stimuli.

The answer in the poor example above is a familiar definition straight out of the textbook, and the distractors are in the teacher's own words.

8.5 Avoid the use of specific determiners. When words such as *never, always*, and only are included in distractors in order to make them false, they serve as flags to the alert student.

Poor Example

To avoid infection after receiving a puncture wound to the hand, you should:
a. Always go to the immunization center to receive a tetanus shot.
b. Be treated with an antibiotic only if the wound is painful.
*c. Ensure that no foreign object has been left in the wound.
d. Never wipe the wound with alcohol unless it is still bleeding.

Better Example

To avoid infection after receiving a puncture wound to the hand, you should always:
a. Go to the immunization center to receive a tetanus shot.
b. Be treated with an antibiotic if the wound is painful.
*c. Ensure that no foreign object has been left in the wound.
d. Wipe the wound with alcohol unless it is still bleeding.

In the poor example above, the underlined word in each of the distractors is a specific determiner. These words have been removed from the better example by rewording both the stem and the distractors.

8.6 Avoid including keywords in the alternatives. When a word or phrase in the stem is also found in one of the alternatives, it tips the student off that the alternative is probably the answer.

Poor Example

When conducting library research in education, which of the following is the best source to use for identifying pertinent journal articles?

a. *A Guide to Sources of Educational Information.*
*b. *Current Index to Journals in Education.*
c. *Resources in Education*
d. *The International Encyclopedia of Education.*

Better Example

When conducting library research in education, which of the following is the best source to use for identifying pertinent journal articles?

a. *A Guide to Sources of Educational Information.*
*b. *Education Index.*
c. *Resources in Education.*
d. *The International Encyclopedia of Education.*

In the poor example above, the underlined word "journal" appears in both the stem and the answer. This clue has been removed from the better example by replacing the answer with another valid answer that does not include the keyword.

Research. Several studies have reported that items are easier when a keyword in the stem is also included in the answer (Haladyna & Downing, 1989b).

8.7 Use plausible distractors. For the student who does not possess the ability being measured by the item, the distractors should look as plausible as the answer. Unrealistic or humorous distractors are nonfunctional and increase the student's chance of guessing the correct answer.

Poor Example

Which of the following artists is known for painting the ceiling of the Sistine Chapel?

a. Warhol.
b. Flintstone.
*c. Michelangelo.
d. Santa Claus.

Better Example

Which of the following artists is known for painting the ceiling of the Sistine Chapel?
a. Botticelli.
b. da Vinci.
*c. Michelangelo.
d. Raphael.

The implausible distractors in the poor example have been replaced by more plausible distractors in the better example.

Plausible distractors may be created in several ways, a few of which are listed below:

- Use common student misconceptions as distractors. The incorrect answers supplied by students to a short answer version of the same item are a good source of material to use in constructing distractors for a multiple-choice item.

- Develop your own distractors, using words that "ring a bell" or that "sound official." Your distractors should be plausible enough to keep the student who has not achieved the objective from detecting them, but not so subtle that they mislead the student who *has* achieved the objective.

9. Avoid the alternatives "all of the above" and "none of the above" (in general).

These two alternatives are frequently used when the teacher writing the item has trouble coming up with a sufficient number of distractors. Such teachers emphasize quantity of distractors over quality. Unfortunately, the use of either of these alternatives tends to reduce the effectiveness of the item, as illustrated in the following table:

Alternative	Use	Weakness
"All of the above"	Answer	Can be identified by noting that two of the other alternatives are correct
	Distractor	Can be eliminated by noting that one of the other alternatives is incorrect
"None of the above"	Answer	Measures the ability to recognize incorrect answers rather than correct answers
	Distractor	Does not appear plausible to some students

Research. While research on the use of "all of the above" is not conclusive, the use of "none of the above" has been found in several studies to decrease item discrimination and test score reliability (Haladyna & Downing, 1989b).

10. Use as many functional distractors as are feasible.

Functional distractors are those chosen by students that *have not* achieved the objective and are ignored by students that *have* achieved the objective. In other words, they have positive discrimination. The following table categorizes distractors according to functionality:

Description	Discrimination	Meaning
Functional	Positive	More non-achievers choose them than achievers
Nonfunctional	Low or none	Achievers and non-achievers choose them equally, or they are rarely chosen at all
Dysfunctional	Negative	More achievers choose them than non-achievers

Whether or not a distractor is functional can be determined through item analysis, a statistical procedure which is discussed in books such as the one by Oosterhof (1990) listed in the bibliography of this booklet.

In general, multiple-choice items contain from two to four distractors. Many teachers assume that the greater the number of distractors in the item, the smaller the chance of guessing the correct answer. This assumption, however, is only true when all of the distractors are functional, and the typical multiple-choice item contains *at least* one nonfunctional distractor. Such

distractors simply fail to distract, and the item would perform just as well if they were omitted entirely.

The solution, therefore, is to only include functional distractors in your items. If you can come up with two or three good ones, avoid the temptation to pad the item with a few poor ones merely to ensure that it has the same number of alternatives as your other items. Such uniformity is artificial, and only serves to lengthen the test without increasing the information provided by the test.

Poor Example

Obsidian is an example of which of the following types of rocks?
*a. Igneous.
b. Metamorphic.
c. Sedimentary.
d. Transparent.
e. None of the above.

Better Example

Obsidian is an example of which of the following types of rocks?
*a. Igneous.
b. Metamorphic.
c. Sedimentary.

Assuming that alternatives *d* and *e* in the poor example above are rarely selected by students, the item is improved by removing these nonfunctional distractors.

Research. Numerous studies have reported that there is little difference in difficulty, discrimination, and test score reliability among items containing two, three, and four distractors (Haladyna & Downing, 1989b).

11. Include one and only one correct or clearly best answer in each item.

When more than one of the alternatives can be successfully defended as being the answer, responding to an item becomes a frustrating game of determining what the teacher had in mind when he or she wrote the item. Such ambiguity is particularly a problem with items of the best-answer variety, where more than one alternative may be correct, but only one alternative should be clearly best. If competent authorities cannot agree on which alternative is clearly best, the item should either be revised or discarded.

In items measuring the student's knowledge of the opinions of others, the name of the individual holding the opinion should be specifically stated.

Poor Example

The United States should adopt a foreign policy based on:
a. A strong army and control of the North American continent.
b. Achieving the best interest of all nations.
c. Isolation from international affairs.
*d. Naval supremacy and undisputed control of the world's sea lanes.

Better Example

According to Alfred T. Mahan, the United States should adopt a foreign policy based on:
a. A strong army and control of the North American continent.
b. Achieving the best interest of all nations.
c. Isolation from international affairs.
*d. Naval supremacy and undisputed control of the world's sea lanes.

While the answer to the poor example above is a matter of debate, the underlined phrase added to the better example clarifies the problem considerably and rules out all of the alternatives except the answer.

12. Present the answer in each of the alternative positions approximately an equal number of times, in a random order.

Many teachers have a tendency to avoid placing the answer in the first or last alternative positions, preferring instead to "bury the answer in the middle." This tendency, however, is not unknown to certain students, who generally select one of the alternatives in the middle if they are unsure of the answer. Also, if there is a noticeable pattern to the positions of the answers from item to item, alert students will take notice and make their selections accordingly. In either case, the unprepared but clever student increases his or her chance of obtaining a higher score.

The easiest method of randomizing the answer position is to arrange the alternatives in some logical order. The following table gives examples of three logical orders. The best order to use for a particular item depends on the nature of the item's alternatives.

Logical order	Example
Numerical	a. 1939 b. 1940 c. 1941 d. 1942
Alphabetical	a. Changing *a* from .01 to .05. b. Decreasing the degrees of freedom. c. Increasing the spread of the exam scores. d. Reducing the size of the treatment effect.
Sequential	a. Heating ice from -100°C to 0°C. b. Melting ice at 0°C. c. Heating water from 0°C to 100°C. d. Evaporating water at 100°C. e. Heating steam from 100°C to 200°C.

Research. Numerous studies indicate that items are easier when this guideline is violated (Haladyna & Downing, 1989b).

13. Lay out the items in a clear and consistent manner.

Well-formatted test items not only make taking the test less confusing and less time consuming for students, they also make grading the test easier, especially when the grading is done by hand. The following suggestions may help you improve the layout of your items.

- Provide clear directions at the beginning of each section of the test.
- Use a vertical format for presenting alternatives.
- Avoid changing pages in the middle of an item.

Poor Example

Your supervisor informs you that three of your fifteen employees have complained to him about your inconsistent methods of supervision. The first thing you should do is a. ask if it is proper for him to allow these employees to go over your head. *b. ask what specific acts have been considered inconsistent. c. explain that you've purposely been inconsistent because of the needs of these three employees. d. offer to attend a supervisory training program.

Better Example

Your supervisor informs you that three of your fifteen employees have complained to him about your inconsistent methods of supervision. The first thing you should do is:

a. Ask if it is proper for him to allow these employees to go over your head.
*b. Ask what specific acts have been considered inconsistent.
c. Explain that you've purposely been inconsistent because of the needs of these three employees.
d. Offer to attend a supervisory training program.

The alternatives are more difficult to locate and compare in the poor example than they are in the better example.

14. Use proper grammar, punctuation, and spelling.

This guideline should be self-evident. Adherence to it reduces ambiguity in the item and encourages students to take your test more seriously.

15. Avoid using unnecessarily difficult vocabulary.

If the vocabulary is somewhat difficult, the item will likely measure reading ability in addition to the achievement of the objective for which the item was written. As a result, poor readers who have achieved the objective may receive scores indicating that they have not.

Use difficult and technical vocabulary only when essential for measuring the objective.

Research. Although very little research has been done on this guideline, one study has reported that simplifying the vocabulary makes the items about 10% easier (Cassels & Johnstone, 1984).

16. Analyze the effectiveness of each item after each administration of the test.

Item analysis is an excellent way to periodically check the effectiveness of your test items. It identifies items that are not functioning well, thus enabling you to revise the items, remove them from your test, or revise your instruction, whichever is appropriate.

For more information on item analysis, refer to a book on educational measurement such as the one by Oosterhof (1990), listed in the bibliography of this booklet.

Bibliography

Albanese, M. A. (1990, April). *Type K and other complex multiple choice items: An analysis of research and item properties.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Boston.

Baker, E. L. (1971). The effects of manipulated item writing constraints on the homogeneity of test items. *Journal of Educational Measurement,* 8, 305-309.

Cassels, J. R. T., & Johnstone, A. H. (1984). The effect of language on student performance on multiple-choice tests in chemistry. *Journal of Chemical Education,* 61, 613-615.

Ebel, R. L., & Frisbie, D. A. (1986). *Essentials of educational measurement* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Frisbie, D. A. (1990, April). *The evolution of the multiple true-false item format.* Paper presented at the Anual Meeting of the National Council on Measurement in Education, Boston.

Grounlund, N. E. (1982). *Constructing achievement tests* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Haladyna, T. M., & Downing, S. M. (1989a). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education,* 2(1), 37-50.

Haladyna, T. M., & Downing, S. M. (1989b). Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education,* 2(1), 51-78.

Hopkins, C. D., & Antes, R. L. (1979). *Classroom testing: construction.* Itasca, IL: F. E. Peacock.

Nitko, A. J. (1983). *Educational tests and measurement: An introduction.* New York: Harcourt Brace Jovanovich.

Oosterhof, A. C. (1990). *Classroom applications of educational measurement.* Columbus, OH: Merrill Publishing.

Ory, J. C. (1983). *Improving your test questions.* Paper identified by the Task Force on Establishing a National Clearinghouse of Materials Developed for Teaching Assistant Training. (ERIC Document Reproduction Service No. ED 285 499)

Osterlind, S. J. (1989). *Constructing test items.* Boston: Kluwer Academic.

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

Zimmerman, B. B., Sudweeks, R. R., Shelley, M.F., & Wood, B. (1990). *How to Prepare Better Tests: Guidelines for University Faculty.* Provo, UT: Brigham Young University Testing Services.

Checklist for Reviewing Multiple-Choice Items

- ☐ Has the item been constructed to assess a single written objective?
- ☐ Is the item based on a specific problem stated clearly in the stem?
- ☐ Does the stem include as much of the item as possible, without including irrelevant material?
- ☐ Is the stem stated in positive form?
- ☐ Are the alternatives worded clearly and concisely?
- ☐ Are the alternatives mutually exclusive?
- ☐ Are the alternatives homogeneous in content?
- ☐ Are the alternatives free from clues as to which response is correct?
- ☐ Have the alternatives “all of the above” and “none of the above” been avoided?
- ☐ Does the item include as many functional distractors as are feasible?
- ☐ Does the item include one and only one correct or clearly best answer?
- ☐ Has the answer been randomly assigned to one of the alternative positions?
- ☐ Is the item laid out in a clear and consistent manner?
- ☐ Are the grammar, punctuation, and spelling correct?
- ☐ Has unnecessarily difficult vocabulary been avoided?
- ☐ If the item has been administered before, has its effectiveness been analyzed?