PREPARING MULTIPLE-CHOICE ITEMS

Guidelines and Instructions
INTRODUCTION

The most commonly used type of test item in admissions, certification, and licensure examinations is the multiple-choice item. The purpose of this manual is to assist test developers and representatives of professional associations in becoming more skilled in writing multiple-choice test items. Preparing multiple-choice items may appear on the surface to be a relatively simple task, given the familiarity with the subject matter, but in reality it requires a great deal of patience and skill.

This manual begins with item-writing guidelines and provides examples using general knowledge as the content in order to facilitate understanding of the concept. Next, basic steps in writing items are presented, along with a template for item development that can be copied and used to develop item sets. Finally, a summary of Bloom's Taxonomy is included along with key words to facilitate the development of more cognitively challenging items.

THE MULTIPLE-CHOICE ITEM

Multiple-choice items consist of two basic parts, the stem and the responses.

The stem is the introductory statement, question, or problem that elicits the correct answer. The responses, sometimes referred to as the answer-set, include the correct and incorrect answers. The correct answer is sometimes referred to as the keyed answer; the incorrect answers are called foils or distractors.

STEM       Who is the scientist most closely associated with the
discovery of the polio vaccine?

RESPONSES  a. Jonas Salk*
b. Louis Pasteur
c. Edward Jenner
d. Robert Koch

Of the four responses in the above example, a is the correct answer and b, c, and d are distractors. One of the key skills of multiple-choice item writing is the ability to write high-quality distractors that effectively discriminate those who have mastered the material being assessed from those who have not.
Items can be written in the form of a question (above) as well as in the form of an incomplete statement.

The polio vaccine was discovered by:
   a. Jonas Salk
   b. Louis Pasteur
   c. Edward Jenner
   d. Robert Koch

Note that the responses for this item are followed by a period because the response serves to complete a sentence.

It is wise to draft an item first as a direct question and revise it into an incomplete sentence if it offers a smoother, less repetitive wording.

**CORRECT VS. BEST-ANSWER ITEMS**

It is important to distinguish between items in which only one response is correct and items in which more than one response may be technically correct, but in which only one response is clearly the best answer.

**Correct-Answer Item**

The capital of Kentucky is
   a. Frankfort.
   b. Lexington.
   c. Louisville.
   d. Ashland.

**Best-Answer Item**

The *major* goal of raising the legal drinking age is to
   a. reduce driving fatalities and injuries.
   b. reduce liquor consumption.
   c. Discourage the habit of drinking by adolescents.
   d. Reduce roadside litter.

In writing best-answer items, care must be taken to word the stem in such a way as to make it clear that only one response is preferred. In the above example, the word *major* achieves this goal.
MULTIPLE-CHOICE ITEM WRITING GUIDELINES

The following guidelines are essential in developing high-quality multiple choice items that neither confuse nor penalize the examinee nor reward the examinee by inadvertently providing clues which make the correct answer obvious. They should be referred to frequently when writing items and when reviewing items written by others.

1. **Express items as precisely, clearly, and simply as possible.**

Do not include any inessential information in the item stem. Stems should ask the question as succinctly as possible.

Unnecessary material reduces the effectiveness of an item by forcing examinees to respond to the irrelevant material and perhaps be distracted by it. For example, the following item:

> In carrying out scientific research, the type of hypothesis which indicates the direction in which the experimenter expects the result to occur once the data has been analyzed is known as a(n)...

could be written

> An hypothesis which indicates the expected results of the study is called a(n)...

Consider also:

**Poor**

In 1850, Adolphe Chattin, then professor of pharmacy in Paris, believed that goiter resulted from an inadequate amount of iodine in the diet. The thyroid is part of which body system?

- a. Endocrine
- b. Nervous
- c. Locomotor
- d. Cardiovascular

**Better**

The thyroid is part of which body system?

- a. Endocrine
- b. Nervous
- c. Locomotor
- d. Cardiovascular
2. **State the problem in the stem.**

The item should be clear enough to provide the examinee with sufficient information to anticipate the type of answer before looking at the responses.

The item

- Multiple-choice items
  - may have several correct answers.
  - consist of a stem and some options.
  - always measure factual details.

does not have a problem or question posed in the stem. The examinee cannot determine the problem on which the item is focused without reading each of the options. The item could be revised to

- The components of a multiple-choice item are a
  - stem and several foils.
  - correct answer and several foils.
  - stem, a correct answer, and some foils.
  - stem and a correct answer.

A student who has been given the objective of recognizing the components of a multiple-choice item will read the stem and know the correct answer.

Also consider the following example:

**Poor**

**Diabetes is:**
- a disorder associated with the body’s absorption of glucose.
- a digestive disorder that afflicts adults.
- a potentially fatal disorder if not detected early
- a disease that can result in blindness.

**Better**

**Diabetes is a chronic disease characterized by**
- the inability of the body to absorb glucose.
- a genetic abnormality resulting in high blood sugar levels.
- an excessive ingestion of sweet foods during infancy.
- an inability to digest amino acids.

Note that in the above example, making the stem more specific clarifies the item significantly.
3. Emphasize general tasks, knowledge, or concepts rather than trivial information or small details.

The item

The product-moment correlation coefficient was developed by
a. John Gosset
b. Sir Ronald Fisher
c. Karl Pearson

might be replaced with

The product-moment correlation coefficient is used to determine the degree of relationship between
a. two dichotomous variables.
b. a dichotomous variable and a continuous variable.
c. two continuous variables.

If an item on the product-moment correlation coefficient is to be included on a test it should concern some basic understanding or skill useful in determining when and how to apply the technique.

Also consider the following item:

Poor

The assassination of President John F. Kennedy occurred on what date?
  a. November 22, 1963
  b. December 4, 1961
  c. February 18, 1961
  d. June 7, 1959

Better

President John F. Kennedy was assassinated
  a. before the passage of the major civil rights legislation of the 1960’s.
  b. before the Bay of Pigs invasion of Cuba.
  c. after the Vietnam war ended.
  d. before the construction of the Berlin Wall.

Note how the better of the two items assesses the sequence of historical events rather than a single date.
4. **Include all qualifications necessary to provide a reasonable basis for responding**

If an item requires an evaluation or judgment, be sure to provide the authority or source of the judgment.

The item

*What is the most effective type of test item?*

might be rewritten

*According to Ebel, the most versatile type of objective item for measuring a variety of educational outcomes is the....*

The second version specifies whose opinion is to be used, narrows the task to consideration of objective items, and focuses on one item characteristic. The first version poses an almost impossible task.

This is especially important when dealing with ethical decisions and decisions regarding interpersonal relationships or controversial subject matter.

**Poor**

The best way to discipline a child is to

a. reward good behavior.

b. use physical punishment for serious offenses.

c. explain to the child why an undesirable behavior displeases you.

d. ignore misbehavior.

**Better**

According to behavioral psychology, the best way to discipline a child is to

a. reward good behavior.

b. use physical punishment for serious offenses.

c. explain to the child why an undesirable behavior displeases you.

d. ignore misbehavior.

Likewise

5. **Do not use “what would you do” or “what do you believe?”**

In items of this kind, the examinee could always defend their response selection as being correct, since each person has a unique frame of reference.
6. **The item stem should be specific enough to pose only one question or problem, and each distractor should be related to that question or problem.**

It is very confusing for the examinee to read an ambiguous stem for which several distractors could be construed as being correct.

**Poor**

Alcoholic beverage consumption is believed to be
a. a tradition in Western cultures.
b. a vice by those with certain religious beliefs.
c. a means to achieve desirable states of relaxation.
d. a major cause of family conflict and disorganization.

**Better**

Data on per capita alcoholic beverage consumption in the United States during the last 10 years suggests what statistical trend?
   a. Increased consumption by females
   b. Decreased consumption by adolescents
   c. Abstinence among all age groups
   d. Abstinence among those over age 65

7. **The correct answer must clearly stand out as the one that experts in the field would recognize as the best response.**

When there is doubt or controversy about the correct response not only is the examinee confused, but the item could be challenged as inappropriate.

**Poor**

The primary cause of the Vietnam war was
   a. Vietnamese nationalism.
   b. French imperialism.
   c. Chinese expansionism.
   d. American imperialism.

**Better**

Which of the following factors most likely sustained the military campaign of the Vietnamese insurgents in the Vietnam war?
   a. Nationalism.
   b. Chinese expansionism.
   c. Soviet militarism.
   d. American expansionism.
8. Avoid negatively stated items – i.e., items that require the examinee to select an incorrect response rather than a correct response.

These types of stems tend to confuse the examinee and tend to make the items less effective in assessing subject matter knowledge and understanding.

Poor

Which of the following authors is not a major American novelist?
   a. George Orwell
   b. Saul Bellow
   c. John Steinbeck
   d. Ernest Hemingway

Better

Which of the following authors is a major British novelist?
   a. George Orwell
   b. Saul Bellow
   c. John Steinbeck
   d. Ernest Hemingway

9. Whenever possible, include in the stem all words that would otherwise have to be repeated in each of the responses.

Poor

The process whereby amniotic fluid is removed from a pregnant woman to test for possible defects
   a. is known as amniocentesis.
   b. is known as cesarean section.
   c. is known as embryonic analysis.
   d. is known as fetal catheterization.

Better

The process whereby amniotic fluid is removed from a pregnant woman to test for possible defects is known as
   a. amniocentesis.
   b. cesarean section.
   c. embryonic analysis.
   d. fetal catheterization.

10. All responses should be grammatically consistent with the tem stem and all response options should be parallel.
Lack of parallelism can sometimes make it possible for examinees to choose the intended answer even when they do not have the skill or knowledge that an item is intended to measure.

Poor

An American city noted for its production of automobiles is
   a. Detroit.
   b. Ohio.
   c. Indiana.
   d. Philadelphia.

Better

An American city noted for its production of automobiles is
   a. Detroit.
   b. Cincinnati.
   c. Indianapolis.
   d. Philadelphia.

11. **Locate and delete irrelevant clues**

Verbal associations and grammatical clues may render an item ineffective.

Poor

A test which may be scored merely by counting the correct responses is an
   ______________________ test.
       a. consistent
       b. objective
       c. stable
       d. valid

Better

A test which may be scored by counting the correct responses is said to be
   a. consistent.
   b. objective.
   c. stable.
   d. valid.

12. **Select diagnostic foils or distractors such as clichés, common misinformation, or logical misinterpretations.**
The purpose of a multiple-choice item is to identify examinees that do not have complete command of the concept or principle involved. To do this, foils or distractors must appear as reasonable as the correct answer to students who have not mastered the material.

For example

A terminal may be defined as
a. a final stage in a computer program.
b. the place where a computer is kept.
c. an input-output device used when much interaction is required.
d. an auxiliary memory unit.

Although option c is keyed as the correct answer, options a and b are derived from the common use of the word “terminal.”

13. Distractors should represent unsafe practices or commonly held misconceptions and should be plausible. Avoid using distractors that even the most uninformed examinee would recognize as being incorrect.

The use of humorous or absurd distractors is not appropriate in standardized test items.

Poor

The person who developed the psychoanalytic theory of human behavior is
a. Sigmund Freud.
b. Babe Ruth.
d. Burt Reynolds.

Better

The person who developed the psychoanalytic theory of human behavior is
a. Sigmund Freud.
b. B. F. Skinner.
c. Carl Rogers
d. John Watson.
14. **Avoid the use of language that is unnecessarily technical or unfamiliar to appropriately trained examinees.**

Always take into consideration the reading level of the examinees.

**Poor**

A biological explanation for the sex difference in human behavior would most likely be supported by research on
- a. fetally androgenized females.
- b. cultural conditioning.
- c. pseudohermaphrodites.
- d. neuroendocrinology.

**Better**

A biological explanation for the sex difference in human behavior would most likely be supported by research on
- a. sex-hormone-level disorders during the fetal stage.
- b. cultural stereotypes of males and females.
- c. people who have both male and female characteristics.
- d. comparisons of male and female athletic performance.

15. **Avoid phrasing the correct answer to a question directly from a textbook.**

Such answers are usually more technical than the distractors and can easily be answered by the examinee with little or no content knowledge.

**Poor**

What is anorexia?
- a. A psychophysiological condition of post-pubital adolescence whereby caloric intake is reduced
- b. A prolonged period of nausea
- c. An adolescent growth disorder
- d. A disorder of the pituitary gland

**Better**

What is anorexia?
- a. Lack or loss of the appetite for foods.
- b. A prolonged period of nausea
- c. An adolescent growth disorder
- d. A disorder of the pituitary gland
16. **Items should not be worded so that they are out of date at the time of testing.**

Poor

In the last presidential election, a vice-presidential candidate who was defeated was
a. Walter Mondale.
b. Robert Dole.
c. Sargent Shriver.
d. George Bush.

Better

In the presidential election of 1980, a vice-presidential candidate who was defeated was
a. Walter Mondale.
b. Robert Dole.
c. Sargent Shriver.
d. George Bush.

17. **Responses should not overlap or include each other so that the examinee could possibly select more than one response and still be technically correct.**

Poor

The right to vote in the United States is granted to individuals of what age?
  a. 18
  b. 16
  c. 17
  d. 19

Better

The right to vote in the United States is granted to individuals beginning at what age?
  a. 18
  b. 16
  c. 17
  d. 19

Note that in the poor example, both a and d could be considered correct.

18. **Do not include the response “none of the above” or “all of the above.”**
It is logically difficult to claim that one of the other responses in the item is either correct or incorrect. For example, when “none of the above” is the correct answer it cannot be determined whether the examinee knows what the correct answer is, only they the wrong answer was identified.

19. **Avoid using absolute modifiers such as always, never, none, and only.**

20. **Avoid providing the examinee with clues that will make it easy to detect the correct answer.**

Do not make the correct answer much shorter, longer, or more technical than the distractors. If this is not possible, try to have two relatively short and two relatively long distractors or two relatively technical and two relatively nontechnical distractors.

Do not put the same key words or descriptive words both in the stem and in the correct answer but not in the distractor.

For example

The major purpose of item analysis is to
a. determine the distribution of test scores.
   b. analyze the pattern of examinee responses.
   c. determine whether the test content was appropriate.
   d. find out if the test is reliable.

Do not make the correct answer clear and concise and the distractors vague and ambiguous.

21. **In certain situations it is useful to base a group of items on one information set. An information set can present a context or relate questions to commonly encountered situations. If this is done, make sure that the information presented in one item does not result in giving away the correct answer for a related item.**
INFORMATION SET

Questions 1 – 3 refer to the following information:

On a biology examination the students obtained the following scores:
90
88
85
76
72

What is the average or mean score for this group of students?
   a. 82.2
   b. 85.0
   c. 79.5
   d. 84.8

If the average score is 82.2, what is the median score?
   a. 85
   b. 88
   c. 76
   d. 90

Note: item 2 is poor because it provides the answer to item 1.

What is the range of scores?
   a. 18 points
   b. 90 points
   c. 72 points
   d. 4 points

22. When possible, avoid the use of gender pronouns such as he, she, his, her, etc.

Neutral items can be constructed by rewording the stem and/or responses to eliminate gender.
Poor

If an elementary school teacher has a child who has been stung by a bee, she should do which of the following things first?
   a. Determine if he is allergic to bee stings.
   b. Remove the stinger from her finger with tweezers.
   c. Attempt to kill the bee.
   d. Bandage the affected area.

Better

Of the following, what is the first thing an elementary school teacher should do to treat a child who has been stung by a bee?
   a. Determine if the child is allergic to bee stings.
   b. Remove the stinger with tweezers.
   c. Attempt to kill the bee.
   d. Bandage the affected area.

23. **An effort should be made to assess higher-level cognitive skills such as application of knowledge and problem solving.**

While it is sometimes difficult to write items of this type, they are needed to assess important components of examinee competence.

Examples:

**Knowledge**

Compared to boys, how much earlier or later is the average start of puberty for girls?
   a. 2 years earlier for girls
   b. 3 years later for girls
   c. 1 year earlier for girls
   d. 2 years later for girls

**Application of Knowledge**

A chaperone at a seventh grade dance would generally observe which of the following conditions?
   a. Girls are taller than boys.
   b. Boys are taller than girls.
   c. Girls and boys are of approximately equal height.
   d. Most of the boys have completed puberty.

**Problem Solving**
If a school principal wanted to arrange for a dance where the boys and girls would be of approximately equal height, this could be best achieved by inviting

a. seventh-grade girls and eighth-grade boys.
b. seventh-grade boys and ninth-grade girls.
c. ninth-grade boys and ninth-grade girls.
d. eighth-grade girls and seventh-grade boys.
STEPS TO WRITING MULTIPLE-CHOICE ITEMS

1. Select a critical topic from your assigned content area

2. Select specific learning objective or domain

3. Determine what the learner is expected to know or do about the topic or within the learning objective/domain

4. Clarify the problem: exactly what do you wish to test for?

5. Consider the cognitive typology of the question: at what level of thinking do you wish the students to engage?

6. Decide what should go into the stem. As a general rule, place in the stem the information of situation that would confront the practitioner.

7. Refer to Bloom's Taxonomy of Objectives in the Cognitive Domain for assistance in selecting verbs

8. List plausible wrong answers such as common misconceptions, outdated information, and potentially harmful errors of omission.

9. Write the item in complete form following the above Guidelines.

10. Write a justification paragraph for the item and reference it to a defensible source.

11. Reference the item to the domain or learning objective it purports to measure.

12. The following Item Construction Sheet is a useful tool and can be photocopied and used when developing items in a workshop setting or when developing rough drafts of items.

13. Finalized items are sent electronically in a Word file to the NRAEF. Each item must contain complete information.
<table>
<thead>
<tr>
<th>Taxonomic Categories and Subcategories</th>
<th>Verbs to use in Objectives</th>
<th>Examples of Appropriate Content in Objectives</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1.0 Knowledge</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.1 Knowledge of specifics</td>
<td>Define, Distinguish, Acquire</td>
<td>Theories, Definitions, Relationships, Examples, Vocabulary words</td>
</tr>
<tr>
<td>1.2 Knowledge of ways and means of dealing with specifics</td>
<td>Identify, Recall, Recognize</td>
<td>Facts, Principles, Causes, Examples, Vocabulary words</td>
</tr>
<tr>
<td>1.3 Knowledge of universals and abstractions</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>2.0 Comprehension</strong></td>
<td>Translate, Estimate, Illustrate, Change, Give in one's own words</td>
<td>Meanings, Samples, Conclusions, Theories, Different views</td>
</tr>
<tr>
<td>2.1 Translation</td>
<td>Conclude, Restate, Explain, Demonstrate, Give in one's own words</td>
<td>Consequences, Implications, Effects, Methods, Different views</td>
</tr>
<tr>
<td>2.2 Interpretation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.3 Extrapolation</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>3.0 Application</strong></td>
<td>Apply, Generalize, Relate, Choose</td>
<td>Principles, Laws, Conclusions, Methods, Abstractions</td>
</tr>
<tr>
<td>3.1 Analysis</td>
<td>Develop, Organize, Restructure, Use</td>
<td>Patterns, Theories, Concepts, Objectives</td>
</tr>
<tr>
<td>3.2 Analysis of elements</td>
<td>Categorize, Distinguish, Identify, Recognize</td>
<td>Statements, Hypotheses, Assumptions, Arguments</td>
</tr>
<tr>
<td>3.3 Analysis of relationships</td>
<td>Deduce, Analyze, Compare</td>
<td></td>
</tr>
<tr>
<td>3.4 Analysis of organizational principles</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>4.0 Synthesis</strong></td>
<td>Document, Write, Tell, Produce, Modify</td>
<td>Positions, Products, Designs, Plans, Solutions</td>
</tr>
<tr>
<td>4.1 Production of a unique idea</td>
<td>Develop, Formulate, Originate, Plan</td>
<td>Hypotheses, Discoveries, Concepts, Objectives</td>
</tr>
<tr>
<td>4.2 Production of a plan</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.3 Derivation of a set of abstract relations</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>5.0 Evaluation</strong></td>
<td>Justify, Assess, Judge, Argue</td>
<td>Opinions, Standards, Consistencies, Precisions, Courses of action</td>
</tr>
<tr>
<td>5.1 Judgments in terms of internal evidence</td>
<td>Decide</td>
<td></td>
</tr>
<tr>
<td>5.2 Judgments in terms of external evidence</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Materials and information from the following organizations are contained in this manual:

Professional Examination Services, New York, NY
Michigan State University Scoring Office, East Lansing, MI
BTLS International, Oak Brook, IL
ITEM CONSTRUCTION SHEET

STEM: ________________________________

____________________________________

____________________________________

____________________________________

RESPONSES:

a. ________________________________

b. ________________________________

c. ________________________________

d. ________________________________

Indicate correct answer with an asterisk (*)

REFERENCE (PUBLICATION AND PAGE NUMBER):

____________________________________

____________________________________

____________________________________

DOMAIN/OBJECTIVE:

____________________________________

____________________________________

ITEM WRITER: __________________________
Angoff Procedure

As part of the Angoff procedure, you will be asked to estimate for each item, how many out of 100 minimally competent servers would correctly answer the item (hereafter referred to as the item rating). Your item ratings will be combined with ratings from other SMEs to obtain an average rating for each item. Finally, a sum of average item ratings for all items in a test will become the cut score for that test. A server whose score on the test equals or exceeds the cut score is considered to have shown mastery of the content in the test.

In providing item ratings, you should pay careful attention to several factors:
- Item difficulty
- Definition of a minimally competent FMP
- The chance factor.

In the present context, item difficulty refers to the level of competency needed to get the item right. The greater the item difficulty, the greater the level of competency needed to answer it correctly.

Defining minimal competence of a server is a difficult but essential part of this approach. Your best professional estimate of expected performance is not based on the entire population of examinees, but of a minimally competent server. SMEs should assess each item from the perspective of a server that just meets performance expectations.

Finally, the chance factor refers to the laws of probability that are generally assumed to operate in the test-taking environment. That is, there is some (non-zero) probability of a correct response on an item for a minimally competent server even when he/she does not know the correct answer to the item. For a multiple-choice item, assuming random guessing over repeated applications, the probability of answering the item with four choices is 25%. Therefore, no item ratings should be less than 25%.
The following survey is designed to evaluate items from the item bank of the Responsible Alcohol Service examination. The NRAEF has previously identified specific tasks required for a server, and exam items were developed based on these tasks. The survey below consists of 50 exam items with the correct item bolded as well as the task upon which the item is based.

Please evaluate the relevance of the content of each item to the task listed above each item. Based on this assessment, each item will be categorized as one of the following:

- **Very relevant** – the item clearly relates to the task.
- **Somewhat relevant** – the item is somewhat related to the task.
- **Not relevant** – the item is not related to the task.