



# **Professional Item Writing**

**for High-Stakes Testing**

**Gregory E. Stone, Ph.D., M.A.**



MetriKs Amerique LLC no longer publishes this handbook. This electronic version is available and may be reprinted without cost.

Professional Item Writing

Copyright (c) 2001 by MetriKs Consulting Limited.

All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording or any information storage and retrieval system, without permission in writing from the publisher.

Printed in the United States of America

**THIS E-BOOK IS DISTRIBUTED WITHOUT CHARGE**

# **MetriKs Professional Item Writing**

## **Table of Contents**

Introduction	4
Pre-Assessment	5
Developing Items	7
Types of Items	9
Answer Options: When is Enough Really Enough?	15
20 Keys to Writing Winning Items	17
Working with Case Scenarios	27
Consider the Reference	29
Taxonomy for Beginners	31
Answers to Pre-Assessment	35



# **MetriKs Professional Item Writing**

The goal of this guide is to provide the learner with the opportunity to become familiar with test development processes, procedures and philosophy. Furthermore, it is hoped that the user will learn the general principles of writing test questions for examinations.

The content presented in this guide depicts health care situations. Your unique profession may be different. Extrapolation to your practice setting is encouraged.

## **Our Philosophy of Certification and Licensure**

Tests should be fair, reflect essential special knowledge, and provide test-takers the best opportunity to demonstrate what they know. Tests should clearly differentiate between candidates with special knowledge from those that cannot demonstrate this ability.

The challenge for the item writer is to provide opportunities for candidates to demonstrate the defined special knowledge of their profession, as described in the curriculum. All test makers are encouraged to follow the American Psychological Association standards of excellence in psychometric and testing principles, both in the area of test development and evaluation. An evaluative process should be developed to ensure that only items meeting quality standards are used to determine whether candidates possess or do not possess the special knowledge associated with their area of practice.



## Pre-Assessment

Please indicate whether each of the following statements is true or false.

1. The more difficult the test, the more fair.
2. The purpose of an examination is to provide a fair evaluation of what an individual knows.
3. A certain percentage of individuals should fail the test.
4. All test items should be referenced from standard texts or journals.
5. Application questions are more difficult than knowledge questions.
6. Examinations should test individual practice.
7. Wider geographical representation in test writers assists in eliminating regional differences.
8. People with high IQs will write the best items.
9. Written examination outcomes measure test-taker competency.





## Developing Items

Any constructed examination must be designed to cover a specified range of material. Test developers should begin by identifying role descriptions for the profession about which they are testing. Most often, this is accomplished by way of a role, job or task analysis. After identifying roles, more specific competency statements may be elaborated. An outline should be developed that identifies the body of knowledge to be tested. Test outlines also generally designate the percentage of items that should appear in each content category.

It is customary for test developers to keep items in pools or banks as they are developed. Through analysis of the bank, deficient areas may be identified and item writing may be targeted to those areas. For instance, if an item bank for an arithmetic test contains 100 addition items, 100 subtraction items, 100 multiplication items and 3 division items, it may be reasonable to target additional item writing in the area of division. The following steps describe how items may be evaluated for possible inclusion in the bank and on an examination.

### Step 1

The item is written.

### Step 2

The item should undergo a basic editorial review during which obvious flaws in spelling, grammar, sentence structure and readability are corrected.

### Step 3

The item should next undergo a review for appropriateness of content, correctness of answer and verification of reference.

### Step 4

The item may be incorporated in the bank and appear on the exam.





## **Types of Items**

There are multiple types of items for inclusion on an examination – three of the most popular are reviewed.

### **True-False Items**

True-False items typically present the test-taker with a statement and ask the respondent to discern whether the statement is true.

These types of items are actually more difficult to write than they first appear. It is often very difficult to write a challenging item that is not esoteric. True-False items also give the test-taker a 50-50 chance of answering correctly, even by guessing. This is likely not the best question type for high-stakes examinations.

### **Matching Items**

Matching items typically present a list of items and a list of definitions. The test-taker is asked to match the item with the definition.

Matching items are not generally used on high-stakes tests. They tend to offer examinees the opportunity to make educated guesses simply by the process of elimination. As such, guessing becomes too great a factor and these items are often poor performers.

### **Multiple-Choice Items**

Multiple-choice items present the test-taker with a question or incomplete statement and a series of 3-5 potential answers.

Multiple-choice items are considered by most test construction experts to be the most effective of all test item types. The fact that these items are more widely used than any other type of item for standardized tests attests to the historic usefulness of the format.



Multiple-choice items are versatile because they can measure a variety of thinking skills. They require a minimum of writing and can be machine scored. Because many multiple-choice items can be answered in a given period, the format allows a wide range of content to be sampled within a single examination.

The major limitation of the format is that they are difficult to construct. The most difficult aspect of writing a multiple-choice item is the development of plausible distractors. Distractor development is a time-consuming task, but critically important to the proper functioning of the item.

Multiple-choice items are formatted in several ways, but all begin with three components: a stem, an answer and a series of “distractors” or incorrect answers.

## **Stem**

The problem or central question presented in the item is called the “stem.” A test of a good stem is that the test-taker should be able to answer the problem without having to read the alternatives presented below it.

## **Answer**

The correct response, validated by reference, is considered the “answer”. It follows the stem of the question and is listed typically in random (alphabetical) order with the incorrect options.

## **Distractor**

A series of incorrect answers or “distractors,” so named because their plausibility should draw away test-takers who are not secure in their knowledge base. Distractors are listed along with the correct answer.



This most basic form of a multiple-choice item is also considered to be the most functional. A *one-best, positive answer* multiple-choice item is considered most useful for at least four reasons. First, the only purpose of the stem is to confront the test-taker with a specific problem or situation. Second, the phrasing required forces the item writer to clearly state the problem. Third, it requires the test-taker to use judgment, reasoning and other types of forward understanding to select not only the correct response but also the *most* correct response. Fourth, and largely a result of the aforementioned reasons, this type of item as a generalization performs best according to most psychometric measures and indicators.

Examples of *One-Best, Positive Multiple Choice Item*:

Stem: A manifestation of chlamydial infection in the neonate is

Answer: 1. eye discharge.

Distractor: 2. nasal duct obstruction.

Distractor: 3. umbilical cord inflammation

Stem: Cigarette smoking during pregnancy places the fetus at increased risk for

Distractor: 1. death in utero.

Answer: 2. low birth weight.

Distractor: 3. neurological defects.

Other forms of multiple-choice items exist and will be discussed presently but each carries with it numerous flaws that increase error problems.



The converse of the one-best, positive multiple-choice item is the *one-worst, negative answer* multiple-choice item. This reversed item typically presents the test-taker with a series of possible answers all are correct except one.

Example of a *One-Worst, Negative Multiple Choice Item*:

Stem: All of the following are considered approved OSHA disinfectants except

Answer: 1. alcohol.

Distractor: 2. bleach.

Distractor: 3. glutaraldehyde 2-3% solution.

In the example, alcohol is the single incorrect response. These items, often called “Except Questions” are very popular. The popularity of these items is explained largely by the simplicity in writing. It is particularly easy to derive multiple correct answers rather than plausible incorrect answers. Unfortunately, these reverse thinking items do not traditionally perform well. They fail psychometrically for two reasons. First, regardless of common commentaries, human beings simply do not think negatively. When presented with an “except” question, the test-taker must first assimilate the item, then internally reverse it, making it a positive statement. This transposition leads to measurement error, not because of a lack of content knowledge, but because of misinterpretation of logic. Error is the test maker’s mortal enemy. The developer must take every precaution to minimize error. Eliminating these items will reduce error.

The items perform poorly for a second reason as well. Many non-psychometricians believe that there is no difference between a negative question and a positive question of identical content. Evidence gathered for decades objectively demonstrates the fallacy of this thought. Positive and negative questions are **not** identical, even when inclusive of identical content. Similarly, when presented with mostly positively phrased items, the introduction of a negatively



phrased item introduces examinee confusion and wastes valuable exam time without justification.

The final type of multiple-choice item is called the **K-Type** or **Complex Multiple Choice** item. These extremely popular items present the test-taker with a problem and multiple solutions.

Example of a **K-Type** or **Complex Multiple Choice Item**:

Stem: Which of the following must be recorded on a patient's clinical record?

1. Date of birth
2. Gender
3. Insurance carrier

- |             |                 |
|-------------|-----------------|
| Distractor: | A. 1, 2 and 3   |
| Answer:     | B. 1 and 2 only |
| Answer:     | C. 2 and 3 only |
| Answer:     | D. 3 only       |

This format was made popular by the *National Board of Medical Examiners* because it was considered closest in process to the natural thought process. Unfortunately, this was too lofty a goal for a test question, and even the venerable institution that gave birth to the item type has largely abandoned it.

It is true that physicians must often juggle several options in making a diagnosis. It is not true however that the physician is offered the specific options ready made. As a thought process, the item sets up an unrealistic situation where the physician would not have to come up with the options by themselves. This destroys the “how we think” argument.

Outside of these theoretical considerations, the items are not widely recommended because they generally produce large amounts of





## Answer Options: When is Enough Really Enough?

Throughout this text, the reader will see examples of items with both three and four answer options. Typical multiple-choice questions include anywhere from three to five answer options. Generally, if the item performs properly (from a statistical analysis) the number of answer options is irrelevant. Test makers are advised to consider the following when deciding on number of answer options.

☼ Fewer options automatically increases the chance of a lucky guess. For instance, a test-taker theoretically has a 25% chance of getting a four-answer option item correct by luck, whereas that same test-taker has a 33% chance of getting a three-answer option item correct by luck.

*In theory this is true, however, unless the passing point is very low the guesser cannot pass an exam by luck. Furthermore, in high stakes situations, test-takers are not randomly guessing because their future depends upon a passing score.*

☼ With fewer options, the test writer should be certain that each distractor is truly functional.

*As long as the wrong answers are good temptations to test-takers who do not know the correct answer, the item will function well with three, four or five options.*

☼ More options do not make good items automatically. It is often very difficult to arrive at plausible wrong answers. Often times, because an answer option may simply be too obviously incorrect, the four or five answer option item becomes *de facto* a three or four answer option item.





## 20 Keys to Writing *Winning* Items

1. Be simple, direct and concise in presentation of the item. Avoid including irrelevant information.

### *Poor Item*

Sepsis is a significant cause of neonatal morbidity and mortality and therefore requires early recognition. One of the earliest signs that can be observed in a septic newborn is

- A. metabolic acidosis.
- B. polycythemia.
- C. temperature instability. \*

### *Revised Item*

An early sign of sepsis in a neonate is

- A. metabolic acidosis.
- B. polycythemia.
- C. temperature instability. \*

2. All response options should parallel the information requested in the stem. For instance, if the stem asks for symptoms of gastrointestinal obstruction all responses should be symptoms. Doing so parallels the *content* of the stem. Additionally, it is important that all responses are grammatically consistent with the stem. Doing so parallels the *structure* of the stem.

*Poor Item*

Medical personnel should be aware that epinephrine may cause the patient to become

- A. nervous.
- B. lower blood pressure.
- C. worsen respiratory condition.

*In the item, distractors B and C are not logically or grammatically consistent with the stem.*

*Revised Item*

Medical personnel should be aware that epinephrine may cause

- A. nervousness.
- B. hypotension.
- C. respiratory problems.

3. Care should be taken when constructing answer options. All distractors should be plausible and homogenous. Distractors should be real terms – making up nonsense terms is frowned upon.

Furthermore, distractors should challenge, not trick the test-taker. For instance, unless pertinent to the question do not arbitrarily introduce new measurement units (i.e. centigrade units in the distractors when using Fahrenheit in the stem).

*An example of a non-real distractor*

Anachoresis is defined as

- A. a new method to electrically sterilize root canals.
- B. localization and fixation of bacteria in an area of inflammation. \*
- C. a Russian princess.



4. To ensure a random order of answers, it is highly recommended that response options be alphabetized by the first letter of the first word in each option.

*Example:*

Spironolactone blocks the action of

- A. estrogen.
- B. GnRH.
- C. testosterone. \*

5. All response options should be approximately the same length. In particular, the correct answer should not typically be longer than the other response options.

*Poor Item:*

Post-operatively a patient who has had orthognathic surgery would most likely be on a diet of

- A. hard foods to strengthen the bone, e.g. peanuts, popcorn, pretzels.
- B. medium foods that would not harm the healing process.
- C. liquids. \*

*Revised Item:*

Post-operatively a patient who has had orthognathic surgery would most likely be on a diet of

- A. hard foods to strengthen the bone, e.g. peanuts, popcorn, pretzels.
- B. medium foods to strengthen the surrounding muscles.
- C. liquids or very soft foods to allow the wound to heal. \*

6. Avoid repeating similar phrases within each response option. The following shows how repetitious phrases can be easily avoided.



*Poor Item:*

During menopause, hormonal changes reflect

- A. decreased estrogen and increased FSH. \*
- B. decreased FSH and LH.
- C. decreased FSH and normal LH levels.

*Revised Item:*

During menopause, hormonal changes reflect decreased

- A. decreased estrogen. \*
- B. increased FSH.
- C. normal FSH.

7. Do not use ambiguous qualifiers such as “often”, “usually”, “occasionally”, “may”, or “about”. Such words require judgments on the part of the test-taker to determine the exact amount to interpret. Similarly, do not use absolute terms such as “always” or “never”. Outside of death, nothing always happens and there are always flukes to disrupt considerations of never. Additionally, avoid using phrases like “associated with” because practically anything may be associated with anything else.

*Poor Item:*

Fibrocystic breast disease is associated with

- A. a single painless lesion.
- B. cyclic breast pain. \*
- C. milk-like discharge.



*Revised Item:*

A sign or symptom of fibrocystic change is

- A. a single painless lesion.
- B. cyclic breast pain. \*
- C. milk-like discharge.

8. Avoid the use of double negatives. In the following example, using the word “not” in the stem and “stop” in one of the responses makes for double negative thinking.

*Poor Item:*

Patients taking metronidazole should not

- A. consume alcohol while taking the drug. \*
- B. stop the drug if vomiting occurs.
- C. take the drug with meals.

*Revised Item:*

Patients taking metronidazole should

- A. avoid consuming alcohol. \*
- B. continue taking the drug if vomiting occurs.
- C. take the drug on an empty stomach.

9. Generally, avoid the use of abbreviations. The exception to this rule is when an abbreviation is truly common parlance, documentable by published textbook references using the abbreviations. Examples include units of measurement such as centimeters, liters, and kVp, and common medical procedures such as EKG, VDRL, PKU.

It is generally accepted that examinees should be presented with a list of common abbreviations that may be used on examinations.



10. Most examinations are nationally or internationally relevant. Therefore, it is particularly important that items address documented standard of care, rather than more regional treatment variations. Similarly, items dealing with controversial treatments or management regimens are typically unacceptable. New or developing trends often cannot be substantiated in readily accessible sources and place the examinees at an unfair disadvantage.

11. Care should be taken when writing items with numbered or lettered options. The options should be in numerical and/or alphabetical order. For consistency of presentation, it is preferable to list numbered options in increasing order.

*Examples of correctly ordered options:*

A. Type I	A. Vitamin A	A. 300-399
B. Type II	B. Vitamin C	B. 400-499
C. Type III	C. Vitamin D	C. 500-599

12. Item options should be independent and mutually exclusive. Options should not overlap. Doing so simply presents a more confused alternative to the K-Type or Complex multiple-choice question.

*Poor Item:*

Symptoms of ectopic pregnancy include

- A. fever, spotting and shoulder pain.
- B. sharp pelvic pain, spotting and shoulder pain. \*
- C. syncope, painless vaginal bleeding and vomiting.



*Revised Item:*

A symptom of ectopic pregnancy is

- A. fever.
- B. spotting. \*
- C. syncope.

*Poor Item:*

The respiratory rate of a normal, healthy adult ranges between

- A. 9-12 breaths per minute.
- B. 12-15 breaths per minute.
- C. 15-18 breaths per minute.
- D. 18-20 breaths per minute.

*If 15 is the normal rate, then both B and C are correct. To revise the item, create ranges that do not overlap.*

13. Avoid verbal expressions, associations or similarities in wording between the stem and correct answer.

*Poor Item:*

Which of the following forms of sterilization kills spores?

- A. Wiping with alcohol
- B. Soaking in detergent solution
- C. Sterilizing in an autoclave \*
- D. Rinsing with chlorhexadine



*Revised Item:*

Which of the following will kill spores?

- A. Alcohol
- B. Detergent solution
- C. Autoclave \*
- D. Chlorhexadine

14. An examination, particularly a high-stakes licensure or certification exam, is not meant as a learning experience. Do not *teach* in the stem, by including extraneous information not necessary for answering the question.

*Poor Item:*

One in nine women today will develop breast cancer. A risk factor for breast cancer is

- A. age over 40. \*
- B. parity greater than 5.
- C. prolonged use of oral contraceptives.

*Revised Item:*

A risk factor for breast cancer is

- A. age over 40. \*
- B. parity greater than 5.
- C. prolonged use of oral contraceptives.

15. Negatively phrased one-worst answer types of questions can be “sneaky”. Using any of the following will reverse the polarity of your item type:

All of the following are true, except ...



The least preferred intervention is ...

Which of the following is not the cause of ...

Which symptom is not a sign of ...

All of the drugs listed have dangerous side effects, with the exception of ...

16. Items should specifically refer to a learning outcome expressed in the content blueprint. If they do not, the examination score cannot be used to make valid inferences about an examinee's ability.

17. Avoid the use of words or phrases that may be considered offensive or harmful to any racial, ethnic, or gender subgroup. Item writers can inadvertently contaminate their writing with stereotypical, prejudicial, or biased language.

18. Present a definite and clear problem or situation in the stem. Examinees should not need to read the complete item text in order to understand what is being asked of them.

*Poor Item:*

When checking and recording vital signs,

- A. the highest pressure exerted on the circulatory system is diastolic.
- B. an example of a blood pressure reading is 120/70. \*
- C. the carotid artery is located on the inner elbow.
- D. the normal adult pulse rate is between 40 and 60 beats per minute.

*To correct this item, a **single** problem should be considered, with appropriate response options following. Examples of a revised stem include:*

Which of the following would be considered a normal blood pressure reading?

Which artery is most commonly used to take a pulse rate?



What is the normal adult pulse rate?

19. Use vocabulary that is as simple as possible, yet appropriate for the purpose of the item and the ability level of the group.

*Poor Item:*

The promiscuous use of sprays, oils and antiseptics in the nose during acute colds is a pernicious practice because it may have a deleterious effect on

- A. the sinuses. \*
- B. red blood cells.
- C. white blood cells.

*Revised Item:*

Frequent use of sprays, oils and antiseptics in the nose during acute colds may result in

- A. sinus infection. \*
- B. olfactory nerve damage.
- C. white blood cell destruction.

20. Avoid the use of “All of the above” and “None of the above”. In practice, item writers tend to make “all of the above” always correct, thereby giving away the answer. Furthermore, the examinee can guess if he or she can at least confirm two of the options as correct. “None of the above” is a problematic choice. If it is the correct answer, but there is another specific correct answer to the question that was not included as an alternative, an examinee can answer the item correctly but still be wrong (i.e. believe that another unlisted, yet incorrect option, is actually true – thus rewarding ignorance with a correct response). If it is used as an incorrect answer, it must be unequivocally so.



## Working with Case Scenarios

The most popular form of the scenario item is a paragraph or two containing a description of a situation or problem with a series of items that pertain to it. The item format may also involve the analysis or interpretation of charts, graphs, photographs, radiographs, lab values, etc.

There are several key considerations to bear in mind when constructing case scenarios.

**Key 1:** The paragraph(s) should describe the process the examinee has to perform or a problem to be solved. It is important that enough information be included in the scenario so that the question may legitimately be answered.

**Key 2:** The situation in the paragraph should require a higher cognitive level of thought to answer, not simply a recall of information. Recall items can be more straightforwardly asked in a traditional multiple-choice item without wasting the examinee's time in reading.

Questions should test the ability of the examinee to synthesize and apply knowledge and information. The scenario should require the examinee to extrapolate information not directly presented in the scenario. In other words, examinees should be required to analyze the information, tap into their own knowledge and experience, and draw a meaningful conclusion.

**Key 3:** The questions should be written in such a manner that they cannot be answered without using *some* information provided in the scenario.

**Key 4:** If the items are to be considered *just another item* in the final scoring of the examination, then the questions should be independent



of one another (non-interlaced). Specifically, if an examinee answers one question incorrectly they should not be precluded from answering the other questions correctly. Scenarios that present hierarchical or otherwise related questions should be reviewed by a trained psychometrician, who will develop an appropriate scoring rubric (framework) for the examination. ***Warning! If you fail to consider this requirement, your test score may be ruled invalid...it is THAT important.***

Example of a non-interlaced scenario:

During the initial history, a woman describes salt craving, diarrhea and general malaise. Her blood pressure is 80/56, her weight is 92 pounds and her skin appears hyperpigmented.

1. The physician would suspect
  - A. chronic adrenal insufficiency. \*
  - B. Cushing's syndrome.
  - C. Hashimoto's disease.
  
2. To rule out other possibilities, a physician would likely order a/an
  - A. ACTH stimulation test. \*
  - B. overnight dexamethasone test.
  - C. thyroid antibody titer.



## Consider the Reference

As suggested, most high-stakes examinations should reflect standard of care. To ensure this standard is followed item writers should consider the following:

### Credibility:

Are the authors recognized in the field?

Is presentation of the material appropriate?

Is the publication current? (Check publication date. Many institutions suggest a reference “shelf life” of 5 years.)

### Availability:

Is the document widely available for purchase or via a library?

Was the document obtained from a national publication and a recognized publisher?

### Content:

Is the information discussed appropriate for the specialty professional to know?

Does it reflect content identified in the test outline?





# Taxonomy for Beginners

Items are typically focused on a specific competency statement within the content outline, but those same items are developed in ways to assess a variety of critical thinking skills relative to the content. Benjamin Bloom, noted educator in the 20<sup>th</sup> century developed a taxonomic system for understanding and explaining how human beings think. Beginning with the simplest thought process and developing towards higher thinking skills. This taxonomy is essential for understanding what skills within the test-taker are being assessed. The taxonomy runs along the following basic hierarchy:

## **Knowledge**

Requires the test-taker to remember or recall facts

## **Comprehension**

Requires the test-taker to relate facts to one another, like simple puzzle pieces.

## **Application**

Requires the test-taker to demonstrate how the remembered facts are employed in concrete situations. For instance, a physician should be able to understand that specific signs and symptoms may relate to specific conditions.

## **Analysis/Synthesis**

Requires the test-taker to go a step further and demonstrate the ability to begin constructing original thought. The test-taker should be able to piece together the facts and come up with a solution to the problem. For instance, that same physician should be able to interpret the lab values and reports presented, and arrive at a differential diagnosis.



## Evaluation

Requires the test-taker to make judgments about the value of material for a specific purpose.

Most commonly, multiple-choice items evaluate knowledge, comprehension and application levels of competency. Analysis/Synthesis level items are difficult to construct but are important to examination developers of higher-level examinations. Evaluation items are very rare on multiple-choice tests because they are nearly impossible to reasonably construct.

The table below presents these taxonomic levels and their item specific construction implications.

<u>Item Level</u>	<u>Stem</u>	<u>Responses</u>
Knowledge & Comprehension	picture procedure list steps terms situation	locate the item list the steps name the procedure define treatment technical name
Application	problem situation equipment topic problem  principle	solution modification operation relate to other topic information needed to solve use in solving a problem
Analysis & Synthesis	concept procedure record  change in a situation	recognize parts identify features relationship of elements what happens to X, Y, Z



## Examples:

### Knowledge

The definitive test for identification of gestational diabetes following a positive screen is a

- A. fasting blood sugar.
- B. glucose tolerance. \*
- C. two-hour postprandial blood sugar.

### Knowledge ➔ **elevated to** ➔ Application

The results of a three-hour glucose tolerance test of a woman at 28-weeks gestation show the following:

Fasting	85 mg/dL
1-hour	190 mg/dL
2-hour	173 mg/dL
3-hour	140 mg/dL

The practitioner should

- A. recognize that these results are normal. \*
- B. repeat the test in one-week after a 100 gm glucose dose.
- C. schedule the patient to come in for diabetes instruction.

### Knowledge

CA-125 levels during pregnancy normally

- A. decrease.
- B. increase. \*
- C. stay the same.

Knowledge ➡ **elevated to** ➡ Application

A 44-year-old woman, six weeks pregnant had CA-125 levels drawn because two family members were recently diagnosed with ovarian cancer. The results showed elevated levels. The practitioner should advise that

- A. a pelvic ultrasound should be performed for further evaluation.
- B. the results are unreliable because pregnancy elevates CA-125 levels. \*
- C. this confirms a diagnosis of ovarian cancer and she should be referred for treatment.

Application ➡ **elevated to** ➡ Analysis/Synthesis

A 44-year-old woman, six weeks pregnant reports that two family members were recently diagnosed with ovarian cancer. An appropriate course of action would be to

- A. monitor blood gases for change.
- B. order a CA-125 draw. \*
- C. refer to oncologist for treatment.



## Answers to Pre-Assessment

1. False

Difficulty and fairness are not necessarily linked.

2. True

This is in fact the only purpose of the examination.

3. False

Percentages of failing or passing candidates are irrelevant. A reasonably set passing standard, using a model such as the Objective Standard Setting system is what counts.

4. True

If test-takers have to read texts with limited availability, they are not given a reasonable chance of learning the information.

5. False

There is not necessarily a relationship between the levels of knowledge tested and the difficulty of the item.

6. False

Tests should be general, covering entire curricula.

7. True

Better representation provides better, multiple perspectives.

8. False

Item writing is part science, but also part art. High performers often tend to write esoteric items.

9. False

Written examinations test knowledge, not skill or competency.







## About the Author

Dr. Stone has been a leader and a pioneer in the field of measurement for over 15 years. Most notable is his development of the Objective Standard Setting model for deriving criterion-referenced passing standards on high-stakes examinations. He has spent his professional career assisting health care related boards to develop quality examination. Dr. Stone holds a BA from Shimer College, attended Oxford University (UK) and Northwestern University, an MA from Loyola University of Chicago and a Ph.D. from The University of Chicago. He is a frequent presenter at national and regional education conferences and has been published widely.

