



NBME[®] ITEM-WRITING GUIDE

**Constructing Written Test Questions
for the Health Sciences**

FEBRUARY 2021



CONSTRUCTING WRITTEN TEST QUESTIONS FOR THE HEALTH SCIENCES

CONTRIBUTORS FROM NBME¹

Melissa S. Billings	Director, Test Development Technology
Kristine DeRuchie	Director, Test Materials Development
Kieran Hussie	Manager, Multimedia Services and Applications
Allison Kulesher	Managing Editor
Jacquelyn Merrell	Managing Editor
Amy Morales	Director, Test Materials Development
Miguel A. Paniagua, MD	Medical Advisor
Jennifer Sherlock	Senior Support Coordinator, Test Materials Development
Kimberly A. Swygert, PhD	Director, Test Development Innovations
Julie Tyson	Senior Editor

Printed copies are not mailed, supplied, distributed, or otherwise made available directly from National Board of Medical Examiners (NBME), except in conjunction with an NBME-sponsored activity (eg, item-writing workshop, meeting or seminar, promotional or otherwise). Additional copies can be obtained by downloading the manual from the NBME website (www.nbme.org). Permission to copy and distribute this document is granted by NBME provided that: (1) the copyright and permission notices appear on all reproductions; (2) use of the document is for noncommercial, educational, and scientific purposes only; and (3) the document is not modified in any way. Any rights not expressly granted herein are reserved by NBME.

¹ Contributing authors for the previous editions of this book also included Susan M. Case, PhD, Steven Haist, MD, MS, and David B. Swanson, PhD.

CONTENTS

06 Section 1: Issues Related to Format and Structure of Test Items

07 Chapter 1: Introduction

Assessment: An Important Component of Instruction

Purposes of Testing

What Material Should Be Tested?

09 Chapter 2: Multiple-Choice Item Formats

15 Chapter 3: Technical Item Flaws

Flaws Related to Irrelevant Difficulty

Flaws That Cue the Testwise Examinee

Summary of Technical Item Flaws

24 Chapter 4: Item Analysis and Interpretation of Results

Analysis of Item Difficulty

Analysis of Item Discrimination

Analysis of Item Options

Comparative Analyses of Test-taker Groups

28 Section 2: Writing One-Best-Answer Items for the Foundational (Basic) and Clinical Sciences

29 Chapter 5: Basic Rules for Writing One-Best-Answer Items

Rule 1: Each item should focus on an important concept or testing point.

Rule 2: Each item should assess application of knowledge, not recall of an isolated fact.

Rule 3: The item lead-in should be focused, closed, and clear; the test-taker should be able to answer the item based on the vignette and lead-in alone.

Rule 4: All options should be homogeneous and plausible to avoid cueing to the correct option.

Rule 5: Each item should be reviewed to identify and remove technical flaws that add irrelevant difficulty or benefit savvy test-takers.



CONTENTS

32 Chapter 6: Testing Application of Foundational (Basic) and Clinical Knowledge

Choosing the Topics to Test

Determining Level of Cognition to Assess

Guidelines for Clinical Vignette Content

The Patient Chart/Table Format

The F-type Format

Guideline(s) for Writing Item Lead-ins

Structuring Items to Fit Task Competencies

51 Chapter 7: Using Media as Part of Clinical Vignettes

Types of Media

Selecting Media

Content Areas Conducive to the Use of Media

Acquiring and Creating Media

Tips for Creating Videos

Media Accessibility

62 Section 3: Appendices

63 Appendix A: A Quick Reference Guide to Approaching Item Writing

64 Appendix B: Sample Lead-ins Based on Task Competencies

79 Appendix C: NBME Retired Item Formats

85 Appendix D: Resources and Additional Reading on Item and Test Development



PREFACE TO THE SIXTH EDITION

The purpose of this manual is, as it has always been, to help faculty members across health professions improve the quality of the multiple-choice items they write for their examinations, and to help them better understand the nature of item development, item analysis, and some basics of psychometrics and measurement. This manual provides a general overview of topics such as item and vignette writing for the foundational and clinical sciences, including technical item flaws, item analysis, and other areas that should provide useful guidance to the health sciences test developer. New content in this edition includes guidance on including patient characteristics in items as well as recent innovations in the use of alternate item types. We anticipate that the primary users of this manual will be faculty members who are teaching health sciences students in basic and clinical science courses as well as in clinical training. The examples and templates focus on undergraduate health sciences education, though the general approach to item writing should be useful for assessing test-takers at other competency levels and across various health care professions.

This manual reflects lessons that the NBME Test Development Services unit has learned in developing items and examinations over the years. During this time, NBME staff members have reviewed countless multiple-choice items and have worked with hundreds of test material development, item writing, and item review committees for United States Medical Licensing Examination® (USMLE®) Step exams, NBME exams, and health profession specialty board exams. In addition, our staff members have been fortunate to conduct item-writing workshops for thousands of item writers, including faculty members at hundreds of national and international health science schools who were interested in developing items for their own examinations. We can personally attest that each committee and workshop attendee has helped us examine our methods, rethink our arguments, and better frame our thoughts regarding how to write high-quality test items. We hope the revised version of this manual continues to be a source of wisdom and encouragement for item writers throughout the health professions world. Finally, we are indebted to our predecessors, on whose broad shoulders we have stood in enhancing this text.

AUGUST 2020

SECTION 1:

ISSUES RELATED
TO FORMAT AND
STRUCTURE OF
TEST ITEMS



SM

CHAPTER 1: INTRODUCTION

ASSESSMENT: AN IMPORTANT COMPONENT OF INSTRUCTION

Assessment, also known as testing, is a critical component of health professions education. When properly used, it can aid in determining the learners' knowledge and skills, based on criteria related to the stated educational goals. A primary purpose of testing is to communicate what you, as the instructor or preceptor and item writer, view as important. Tests are a powerful motivator, and your test-takers or students will learn the educational concepts they believe you value. Assessment also helps to fill instructional gaps by motivating students to seek educational resources and opportunities beyond course work. This outcome of testing is especially important in clinical learning environments, in which the experienced curriculum may vary from student to student, depending on factors such as the setting and the flow of patients. This outcome may also be important in some basic (foundational) science settings, in which the educational experiences may also vary from student to student. As students progress toward competence or even excellence, they are aided by ongoing feedback from their instructors or preceptors. Tests are often an important and useful component of that feedback in activating further learning.

Because tests have such a powerful influence on student learning, it is important to develop tests that properly align with educational goals and objectives. This manual focuses on the process of writing high-quality multiple-choice questions (MCQs), or items, that can be used to assess a wide variety of clinical knowledge and skills within the framework of the basic and clinical sciences.

Two issues that are of concern when developing and constructing quality MCQ-based exams are content sampling and psychometric performance.

Issues of Content Sampling

The purpose of any assessment is to permit inferences to be drawn concerning the ability (knowledge, skills) of test-takers or examinees. Inferences are defined here as decisions, judgments, or conclusions that extend beyond the set of items included in the exam into the larger domain from which the items were sampled. Performance on the test provides a basis for estimating achievement in the broader domain of interest, and this broader domain should be made explicit with statements about the inferences to be made from the test.

The first decision to be made involves the content to be sampled on the test; content decisions will drive the number and topic areas of the MCQs to be developed. The amount of attention given to evaluating a content area should reflect its relative importance, and it is often impractical to cover all testing topics in equal lengths. Tests are point-in-time measurements that take a specific and limited amount of time; if one content area contains many items, there is less testing time for other content areas. The nature of the test determines the extent to which the estimate of achievement is reproducible (aka reliable or generalizable) and accurate (provides sufficient validity evidence to support the intent). If the test questions are not adequately representative of the broader domain of interest (eg, if a test of competence in general medical practice includes only cardiovascular-related content), the test results may be biased and may not provide a good basis for estimating achievement in the full domain of interest. If the overall test length is too short, the scores may not be sufficiently precise or reliable to ensure they are a good representation of true proficiency. In order to generate a reproducible score, the item writer needs to sample content broadly (ie, typically 100 or more MCQs for written assessments).

Issues of Psychometric Performance

The process of evaluating the psychometric characteristics of an assessment and weighting their relative importance is determined by the purpose of the test and the decisions that will be made based on the results. For tests with higher stakes, such as those used for promotion or graduation decisions, those used for course grades, or those used in isolation for decisions, the scores must be reasonably reproducible (as demonstrated by high reliability) and evidence should be presented to demonstrate the accuracy of the test (eg, showing how content outlines for the test match the inferences to be made). For tests with lower stakes, such as those for which the score is but one element of the decision-making process, the amount of required psychometric evidence is less, but attention should be paid to evidence of test reliability and validity of score use nonetheless (see Figure 1 in Chapter 6 for more information).

PURPOSES OF TESTING

- ▶ Inform students about material that is important
- ▶ Motivate students to study
- ▶ Identify areas of deficiency, in need of remediation, or further learning
- ▶ Determine final grades or make promotion decisions
- ▶ Identify areas in which instruction can be improved

WHAT MATERIAL SHOULD BE TESTED?

- ▶ Exam content should align with course or clinical experience objectives
- ▶ Predetermined important topics should be weighted more heavily than less important topics
- ▶ The testing time devoted to each topic should reflect the relative (predetermined) importance of the topic
- ▶ The breadth of items should be representative of the instructional goals and objectives (curricular alignment)

NOTES

CHAPTER 2: MULTIPLE-CHOICE ITEM FORMATS

One of the most crucial aspects of a multiple-choice test item or question (MCQ) is its type or structure. Different item types can be used for different topic areas, and each item type carries with it advantages and disadvantages. A critical aspect to consider when choosing an item type is the inclusion of potential flaws that might benefit the savvy test-taker or introduce irrelevant difficulty. This chapter covers the basics of several multiple-choice item formats and introduces some potential flaws that are common to specific formats, while Chapter 3 will discuss specific item flaws in more detail.

ONE-BEST-ANSWER VS. TRUE-FALSE ITEMS

The universe of multiple-choice items can be divided into two families:

- ▶ Items that require test-takers to indicate a single, most accurate response (one-best-answer)
- ▶ Items that require test-takers to indicate all responses that are appropriate (true-false)

NBME has used multiple item formats within each family in the past, listed below by designating letter.

One-best-answer item formats that require test-takers to select the single best response:

- ▶ A-type (4 or more options, single items or sets)
- ▶ F-type (2 to 3 items grouped into a set around specific content or case scenario where test-takers **cannot** return to previously seen items in the set)
- ▶ G-type (2 or 3 items grouped into a set around specific content where test-takers **can** return to previously seen items in the set)

True-false item formats that require test-takers select some set of options that are true:

- ▶ C-type (A/B/Both/Neither response items)
- ▶ K-type (complex true-false items)
- ▶ X-type (simple true-false items)

The letters used to label the item formats hold no intrinsic meaning; letters were assigned more or less sequentially to new item formats as they were developed. For an extended list of item types formerly used by NBME, ordered by their designated letters, see Appendix C: NBME Retired Item Formats.

The One-Best-Answer Family

In contrast to true-false questions, one-best-answer questions are designed to make explicit that only one option is to be selected. These items are the most widely used multiple-choice item format. They consist of a stem, which most often includes a vignette (eg, a clinical case presentation) and a lead-in question, followed by a series of option choices, with one correct answer and anywhere from three to seven distractors. The incorrect option choices should be directly related to the lead-in and be homogeneous with the correct answer. This item describes a situation (in this instance, a patient scenario) and asks the test-taker to indicate the most likely cause of the problem.

Sample Stem (Vignette & Lead-in) With Option Set



VIGNETTE:

A 32-year-old man comes to the office because of a 4-day history of progressive weakness in his extremities. He has been healthy except for an upper respiratory tract infection 10 days ago. His temperature is 37.8°C (100.0°F), pulse is 94/min, respirations are 42/min and shallow, and blood pressure is 130/80 mm Hg. He has symmetric weakness of both sides of the face and the proximal and distal muscles of the extremities. Sensation is intact. No deep tendon reflexes can be elicited. Babinski sign is absent.

LEAD-IN:

Which of the following is the most likely diagnosis?

OPTION SET:

- A. Acute disseminated encephalomyelitis
- B. Guillain-Barré syndrome*
- C. Myasthenia gravis
- D. Poliomyelitis
- E. Polymyositis

Note that the incorrect options are not wholly wrong. The options can be diagrammed as follows:



Even though the incorrect options are not completely wrong, they are less correct than the “keyed answer” (indicated by the asterisk in the option set). The test-taker is instructed to select the “most likely diagnosis.” Experts would all agree that the most likely diagnosis is B; they would also agree that the other diagnoses are somewhat likely, but less likely than B. As long as the options can be laid out on a single continuum, in this case from “Least Likely Diagnosis” to “Most Likely Diagnosis,” distractors in one-best-answer items do not have to be totally wrong.

“Cover-the-Options” Rule

This leads us to another important guideline for writing good one-best-answer items—the “cover-the-options” rule. If a lead-in is properly focused, a test-taker should usually be able to read the vignette and lead-in, cover the options, and guess the correct answer without seeing the option set. For example, in this next item, after reading the lead-in, the test-taker should be able to answer the item without seeing the options. When writing items, covering the options and attempting to answer the item is a good way to check whether this rule has been followed.

Example of “Cover-the-Options” Rule

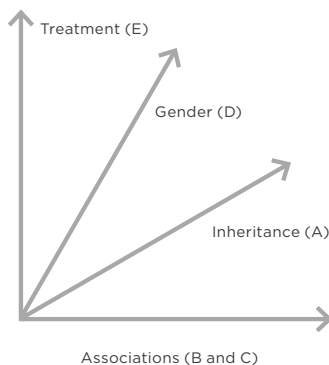


A 58-year-old man comes to the office because of pain in the right knee for the past 3 days. He has a history of type 2 diabetes mellitus, hypertension, and hyperlipidemia controlled with daily glyburide, lisinopril, and atorvastatin. There is no family history of similar disorders. On physical examination, the knee is swollen, passive motion produces pain, and ballottement discloses an effusion. Synovial fluid is cloudy and contains positive birefringent crystals and no bacteria. X-ray shows chondrocalcinosis. Which of the following is the most appropriate pharmacotherapy?

- A. Allopurinol
- B. Betamethasone
- C. Ibuprofen*
- D. Infliximab
- E. Levofloxacin

Homogeneous Options

Along with a focused lead-in, a good item will have a keyed answer and distractors that are homogeneous. They all directly address the lead-in in the same manner and can be rank ordered along a single dimension. The one-best-answer example below is a flawed item that can occur when options are not listed on a single dimension. After reading the lead-in, the test-taker has only the vaguest idea what the question is about. In order to determine the “best” answer, the test-taker must decide whether “it occurs frequently in women” is more or less true than “it is seldom associated with acute pain in a joint.” The diagram of these options might look like the figure to the left of the sample item below. The options are heterogeneous and deal with miscellaneous facts; they cannot be rank ordered from least to most true along a single dimension. Although this item appears to assess knowledge of several different points, its inherent flaws preclude this. The item by itself is not clear; the item cannot be answered without looking at the options.



Which of the following is true about pseudogout?

- A. It is clearly hereditary in most cases
- B. It is seldom associated with acute pain in a joint
- C. It may be associated with a finding of chondrocalcinosis
- D. It occurs frequently in women
- E. It responds well to treatment with allopurinol



The True-False Family

True-false items require test-takers to select all the options that are “true,” which could be anywhere from one to all of the listed options. In answering these items, the test-taker must decide where to make the cutoff and determine to what extent a response must be “true” in order to be keyed as “true.” While this task requires additional judgment beyond what is required to select the true answer(s), that additional judgment may be unrelated to clinical expertise or knowledge. Too often, test-takers have to guess what the item writer had in mind because the options are not either completely true or completely false.

Sample of Acceptable True-False Item

Which of the following are X-linked recessive conditions?

1. Cystic fibrosis
2. Duchenne muscular dystrophy
3. Hemophilia A (classic hemophilia)
4. Tay-Sachs disease



This item is an example of a reasonably acceptable true-false item from a structural perspective. Note that the stem is clear, and the options are absolutely true or false with no ambiguity. Following tradition, for true-false items, the options are numbered. Options should be homogeneous (all are conditions), clearly worded, and of similar length, and the question should be closed and focused.

The options can be diagrammed as follows.



Sample of Flawed True-False Item

True statements about cystic fibrosis (CF) include:

1. CF is an autosomal recessive disease
2. Patients with CF usually live into adulthood
3. Males with CF are sterile
4. The incidence of CF is 1:2000



This item demonstrates a commonly seen flaw for true-false items that often occurs when options are not homogeneous and vaguely worded. Options 2, 3, and 4 cannot be judged as absolutely true or false, because a group of content experts would not necessarily agree on the answers. For example, for option 4, experts would demand more information to determine incidence: Is this in the United States? Is this among all ethnic groups? Similar issues arise with options 2 and 3, whereas option 1 is clear. Revision of this item would most likely include editing options 2, 3, and 4, to be statements of fact like option 1, and revising the question itself to be closed.

Sample of Flawed True-False Item



In children, ventricular septal defects are associated with:

1. cyanosis
2. pulmonary hypertension
3. systolic murmur
4. tetralogy of Fallot

The problems with this true-false item are more subtle. The difficulty is that the test-taker has to make assumptions about the severity of the disease, the age of the patient, and whether or not the disease has been treated. This is due in part to the vagueness in the question itself (“associated with”). Different assumptions lead to different answers, even among experts. Revising this question would require adding additional text, perhaps a lot of it, in order to allow the test-taker to judge the options as wholly true or wholly false.

General Rules for True-False Items

Because test-takers are required to select all the options that are “true,” true-false items must satisfy the following rules:

- ▶ Item and option text must be clear and unambiguous. Avoid imprecise phrases such as “is associated with” or “is useful for” or “is important”; words that provide cueing such as “may” or “could be”; and vague terms such as “usually” or “frequently.”
- ▶ The lead-in should be closed and focused.
- ▶ Options must be absolutely true or false; no shades of gray are permissible.
- ▶ Options should be homogeneous so that they can be judged as entirely true or entirely false on a single dimension.

Challenges with Using True-False Items

We recommend avoiding true-false questions if possible. Although many item writers believe true-false items are easier to write than one-best-answer items, this type can often be more problematic. The writer may have something particular in mind when writing the item, but careful review subsequently reveals subtle difficulties that were not apparent to the item author. Often the distinction between “true” and “false” is not clear, and it is not uncommon for subsequent reviewers to alter the answer key. As a result, reviewers tend to rewrite or discard true-false items far more frequently than items written in other formats. Some ambiguities can be easily clarified, but others cannot. In addition, to avoid ambiguity true-false questions often test on recall of an isolated fact, which we recommend avoiding.

ONE-BEST-ANSWER QUESTIONS ARE BETTER AT ASSESSING TEST TAKERS' JUDGMENT, SYNTHESIS, AND APPLICATION OF KNOWLEDGE.

CHAPTER 3: TECHNICAL ITEM FLAWS

Good content and good structure contribute to the quality of an item. However, quality can be impacted negatively by the inclusion of technical item flaws. There are two kinds of technical item flaws:

1. A flaw that adds irrelevant difficulty to the item can confuse all test-takers. These flaws make the item challenging for reasons unrelated to the testing objective/point of the item and can add construct-irrelevant variance to the final test score.
2. A flaw that cues the more savvy and confident test-takers (aka the “testwise”) and aids them in guessing the right answer. These flaws related to “testwiseness” make it easier for some students to answer the item correctly based on their test-taking skills alone, without necessarily knowing the content.

The item writer’s goal is to develop and structure items to eliminate both types of flaws as much as possible, in order to create a test that ensures a level playing field for all test-takers. A test-taker’s probability of answering an item correctly should be determined by his or her amount of expertise on the topic being assessed; ideally, that probability will not decrease due to a suboptimally written item and will not increase due to test-taking strategies.

FLAWS RELATED TO IRRELEVANT DIFFICULTY

Long or Complex Options

The item below has several flaws. The vignette contains extraneous information, and in fact, the vignette is not needed to answer the item. More importantly, the options themselves are overly long and complicated. The number of words in each option increases the reading load, which can shift the construct that is being measured from content knowledge to reading speed. Please note that this flaw relates only to options. There are many well-constructed test items that include a long vignette, and decisions about vignette length should be made in accordance with the testing point of the item. If the purpose of the item is to assess whether or not the student can interpret and synthesize information to determine, for example, the most likely diagnosis for a patient, then it is appropriate for the vignette to include a fairly complete description of the situation.

Example of Item with Long, Complex Options

Peer review committees in HMOs may move to take action against a physician’s credentials to care for participants of the HMO. There is an associated requirement to ensure that the physician receives due process in the course of these activities. Due process must include which of the following?

- A. Notice, an impartial forum, counsel, and a chance to hear and confront evidence
- B. Proper notice, a tribunal empowered to make the decision, a chance to confront witnesses, and a chance to present evidence in defense
- C. Reasonable and timely notice, an impartial panel empowered to make a decision, a chance to hear evidence and to confront witnesses, and the ability to present evidence in defense



Numeric Data Presented Inconsistently

When numeric options are used, the options should be listed in numeric order and in a single format (ie, as terms or ranges). Confusion can occur when formats are mixed or when options are listed in an illogical order. In this example, options A, B, and C are expressed as ranges, whereas options D and E are specific percentages. All options should be expressed as ranges or as specific percentages; mixing them is ill-advised. In addition, the range for option C includes options D and E, which almost certainly rules out options D and E as correct answers for the testwise examinee.

Example of Item with Inconsistent Numeric Data



After a second episode of infection, which of the following is the likelihood that a woman is infertile?

- A. Less than 20%
- B. 20 to 30%
- C. Greater than 50%
- D. 75%
- E. 90%

Vague Terms

Vague frequency terms in the options (such as “often” or “usually”) are not consistently defined or interpreted by the readers, and sometimes not even by experts. Different interpretations of these terms can lead to multiple correct answers or a set of options that cannot be rank ordered in terms of correctness.

Example of Item with Vague Terms



Severe obesity in early adolescence:

- A. has a 75% chance of clearing spontaneously
- B. often is related to endocrine disorders
- C. shows a poor prognosis
- D. usually responds dramatically to dietary regimens
- E. usually responds to pharmacotherapy and intensive psychotherapy

“None of the Above”

The phrase “None of the above” is problematic in items for which judgment is involved and the options are not absolutely true or false. If the correct response is intended to be one of the other listed options, knowledgeable students are faced with a dilemma because they have to decide between the option that the item writer has intended as correct and an option that encompasses everything not listed in the option set. Test-takers can often intuit an option that is more correct than the item writer intended to be correct, which would lead them to use the more expansive option. Use of “None of the above” essentially turns the item into a true-false item; each option has to be evaluated as more or less true than the universe of unlisted options. It is often possible to fix such items by replacing “None of the above” with an option that is more specific. In this example, which asks a test-taker to specify the most appropriate pharmacotherapy, option E, “None of the above” should be replaced by “No pharmacotherapy is indicated at this time,” to eliminate any ambiguity while still requiring the test-taker to commit to a management decision.

Example of Item with “None of the Above”



A 3-day-old male newborn is brought to the office by his parents because his crying has increased during the past night compared with his first 2 days of life. The parents have been unable to calm the newborn during the past 2 hours. The newborn also has had mild shaking of his hands and legs during the past 4 hours. He was delivered at 38 weeks' gestation via uncomplicated spontaneous vaginal delivery. His mother, gravida 2, para 2, is age 19 years. She has a history of dysthymia for which she took escitalopram during pregnancy. The newborn takes no medications. He is at the 50th percentile for length, weight, and head circumference. Temperature is 37.2°C (98.9°F), pulse is 155/min, respirations are 35/min, and blood pressure is 84/50 mm Hg. Pulse oximetry on room air shows an oxygen saturation of 100%. The newborn has a high-pitched cry and is inconsolable with swaddling. He has tremors of his hands and feet with crying. Moro reflex is present. Which of the following is the most appropriate pharmacotherapy?

- A. Citalopram
- B. Lorazepam
- C. Morphine
- D. Naloxone
- E. None of the above*

Nonparallel Options

The next item illustrates a common flaw in which the options are not only too long but the structure of each option is different, both of which add to the reading time. Generally, this flaw can be corrected by careful editing to ensure that the options all have the same format and the same structure. In this particular item, the lead-in can be changed to “Which of the following is the most likely reason no conclusion can be drawn from these results?” Each option can then be edited to fit a logical and parallel answer to the lead-in.

Example of Item with Nonparallel Options



In a vaccine trial, 200 two-year-old boys were given a vaccine against a certain disease and then monitored for 5 years for occurrence of the disease. Of this group, 85% never contracted the disease. Which of the following statements concerning these results is correct?

- A. The number of cases (ie, 30 cases over 5 years) is too small for statistically meaningful conclusions
- B. Vaccine efficacy (%) is calculated as $85 - 15/100$
- C. No conclusions can be drawn because the trial involved only boys
- D. No conclusion can be drawn since no follow-up was done with nonvaccinated children

Complicated Stems

This item, as written, requires that the test-taker (a) understands the concepts of genetics that are represented and (b) is able to rank order Roman numerals (the second of which is an irrelevant and unnecessarily difficult addition to the goal of the item). This item should be rewritten to focus on a single karyotype, such as the greatest risk, with the karyotypes arranged in the options themselves, so that the test-taker who understands the order of risk of occurrence can more easily identify the correct answer.

Example of Item with Complicated Stem



Arrange the parents of the following children with Down syndrome in order of highest to lowest risk of recurrence. Assume that the maternal age in all cases is 22 years and that a subsequent pregnancy occurs within 5 years. The karyotypes of the daughters are:

I: 46,XX,-14,+T(14q21q)pat

II: 46,XX,-14,+T(14q21q)de novo

III: 46,XX,-14,+T(14q21q)mat

IV: 46,XX,-21,+T(14q21q)pat

V: 47,XX,-21,+T(21q21q) (parents not karyotyped)

- A. III, IV, I, V, II
- B. IV, III, V, I, II
- C. III, I, IV, V, II
- D. IV, III, I, V, II
- E. III, IV, I, II, V

Suggested Revision



Five couples come to the office for counseling prior to conception. Each couple has one child with Down syndrome. The karyotypes of each of the children are shown. The parents of the child with which of the following karyotypes have the greatest risk for recurrence of Down syndrome in their next pregnancy?

- A. 46,XX,-14,+T(14q21q)pat
- B. 46,XX,-14,+T(14q21q)de novo
- C. 46,XX,-14,+T(14q21q)mat
- D. 46,XX,-21,+T(21q21q)pat
- E. 47,XX,-21,+T(21q21q) (parents not karyotyped)

Negatively Phrased Lead-ins

A negative phrasing in the lead-in asks the test-taker to find the least accurate option, with the rest being accurate, rather than to find the most accurate option. If most of the items on a test are positively phrased, the inclusion of a negatively phrased item carries the risk that the test-taker will miss the word “except,” even when it is set in bold and/or capitalized.

Example of Item With Negatively Phrased Lead-in



Each of the following statements about cholesterol is true EXCEPT:

- A. cholesterol contains numerous fatty acids
- B. cholesterol is not present in any foods of plant origin
- C. cholesterol is required in many complex bodily functions
- D. endogenous cholesterol is produced within the body

FLAWS THAT CUE THE TESTWISE EXAMINEE

Grammatical Cues

This flaw exists when an option does not follow grammatically from the lead-in. In this example, testwise students can eliminate B, C, D, and E as possible correct answers because they do not grammatically or logically follow the lead-in. This flaw can happen when an item writer focuses more attention on writing the correct answer than on the distractors, leading to the potential for grammatical errors. To avoid this flaw, read each option immediately following the stem to ensure that the language is a good fit. Another way to avoid the flaw is to always use closed lead-ins.

Example of Item with Grammatical Cue



A 12-year-old girl is brought to the office because of chest pain. She has recently experienced an upper respiratory infection with frequent coughing. Temperature is 37.2°C (99.0°F), pulse is 120/min, respirations are 22/min, and blood pressure is 95/65 mm Hg. Pulse oximetry on room air shows an oxygen saturation of 99%. Physical examination shows tenderness to palpation over her costochondral joints on the left. Auscultation of the lungs discloses diffuse end-expiratory wheezes bilaterally. Her diagnosis is most likely to be an:

- A. asthma attack*
- B. costochondritis
- C. pleurisy
- D. rib fracture secondary to coughing
- E. viral pneumonitis

Grouped or Collectively Exhaustive Options

This flaw exists when a savvy student can identify a subset of options that cover all possible outcomes (are collectively exhaustive) and rule out the options not in that subset. In this item, options A, B, and D are exhaustive—urine potassium can only increase, decrease, or not change—and thus one of these three options must be the correct answer. A less testwise student might spend time considering C and E. Often, item writers add options like C and E only because they want to have a total of five options, but it is not an improvement of the item to add options that have no merit. The item writer should be able to rank order each option on the same dimension, and no subset of options should include all possible outcomes.

Example of Item With Collectively Exhaustive Options



Administration of furosemide results in:

- A. a decrease in urine potassium
- B. an increase in urine potassium
- C. improved glucose control in patients with type 2 diabetes mellitus
- D. no change in urine potassium
- E. requires decreasing the dose with renal failure

Absolute Terms

In this item, options A, B, and E contain terms that are less absolute than those in options C and D. The testwise student will eliminate options C and D as possibilities because they are less likely to be true than something stated less absolutely, and so this item is flawed with the inclusion of those terms. This flaw tends to arise when verbs are included in the options rather than in the lead-in. Focusing the lead-in, placing the verb in the lead-in, and shortening the options are possible ways to correct this flaw.

Example of Item With Absolute Terms



In patients with advanced dementia, Alzheimer type, the memory defect:

- A. can be treated adequately with phosphatidylcholine (lecithin)
- B. could be a sequela of early parkinsonism
- C. is never seen in patients with neurofibrillary tangles at autopsy
- D. is never severe
- E. possibly involves the cholinergic system

Correct Option Stands Out

In this item, the correct answer, option A, is longer than the other options, and is the only “double” option, containing two components. This flaw is another potential outcome when item writers pay more attention to constructing the correct answer than the distractors. This results when item writers likely create the correct answer first and then write the incorrect distractors. In addition, item writers are often teachers and they will construct long correct answers that include additional instructional material, parenthetical information, caveats, and so on. This flaw can be avoided by reviewing the entire option set for length, ensuring the level of detail is consistent across options, and removing language that is purely for instructional purposes only.

Example of Item with Correct Option that Stands Out



Secondary gain is:

- A. a complication of a variety of illnesses and tends to prolong many (>3) of them*
- B. a frequent problem in obsessive-compulsive disorder
- C. never seen in organic brain damage
- D. synonymous with malingering

Word Repetition (“Clang Clues”)

This flaw arises when language used in the stem is repeated in the correct answer. Here, the word “unreal” in the vignette can clue test-takers to the fact that the correct answer, “derealization,” is the only option that also includes the word “real.” The same flaw can appear even if a word is repeated only in an etymological sense, such as when a stem mentions bone pain and the correct answer begins with the prefix “osteo-.” Item writers should scan the options and item stem to check for this word or phrase repetition.

Example of Item with Word Repetition



A 58-year-old man with a history of heavy alcohol use and previous psychiatric hospitalization is confused and agitated. He speaks of experiencing the world as unreal. Which of the following best describes this symptom?

- A. Depersonalization
- B. Derailment
- C. Derealization*
- D. Focal memory deficit
- E. Signal anxiety

Convergence

This item flaw might be less obvious than the others, but it occurs frequently and is worth noting. The underlying flaw is that the correct answer is the option that has the most in common with the other options, and thus the testwise test-taker can converge on the right answer just by counting the number of times certain terms appear. In this example, the testwise test-taker would eliminate “anionic form” as unlikely because “anionic form” appears only once; that test-taker would also exclude “outside the nerve membrane” because “outside” appears less frequently than “inside.” The test-taker would then have narrowed the options to B and D. Since three of the five options involve a charge, the testwise test-taker would then select option B, which is in fact the correct answer. This flaw can also occur without being directly reflected in the language; for example, if an item is asking which pharmacotherapy is most effective, and three of the five options are in one class of drugs, the savvy test-taker may rule out the other two as less likely. This flaw occurs when item writers start with the correct answer and write the distractors as permutations of the correct answer. The correct answer will then be more likely to have elements in common with the rest of the options, and the incorrect answers are more likely to be outliers. A useful check is to review all options and see if words or terms are repeated across options.

Example of Item with Convergence



Local anesthetics are most effective in the:

- A. anionic form, acting from inside the nerve membrane
- B. cationic form, acting from inside the nerve membrane*
- C. cationic form, acting from outside the nerve membrane
- D. uncharged form, acting from inside the nerve membrane
- E. uncharged form, acting from outside the nerve membrane

NOTES

SUMMARY OF TECHNICAL ITEM FLAWS

Issues Related to Irrelevant Difficulty	
FLAWS	SOLUTIONS
Long, complex options	<ul style="list-style-type: none"> ▶ Put common text in stem. ▶ Use parallel construction in options. ▶ Shorten options.
Tricky, unnecessarily complicated stems	<ul style="list-style-type: none"> ▶ Include content that is necessary to answer the question or to make distractors attractive. ▶ Avoid teaching statements.
Inconsistent use of numeric data	<ul style="list-style-type: none"> ▶ Avoid overlapping options. ▶ Ask for minimum or maximum value to avoid multiple correct answers.
Vague terms	<ul style="list-style-type: none"> ▶ Avoid frequency terms, like usually and often. Such terms are interpreted differently by different people.
“None of the above” option	<ul style="list-style-type: none"> ▶ Replace “None of the above” with specific action (eg, No intervention needed).
Nonparallel options	<ul style="list-style-type: none"> ▶ Edit options to be parallel in grammatical form and structure.
Negatively structured stem (eg, “Each of the following EXCEPT”)	<ul style="list-style-type: none"> ▶ Revise lead-in to have a positive structure. ▶ If possible, use correct options to create a scenario.

Cues to the Testwise Examinee	
FLAWS	SOLUTIONS
Collectively exhaustive options (subset of options cover all possibilities)	<ul style="list-style-type: none"> ▶ Replace at least one option in subset. ▶ When revising, avoid creating option pair.
Absolute terms (“always,” “never”) in options	<ul style="list-style-type: none"> ▶ Eliminate absolute terms. ▶ Use focused lead-in and short homogeneous options.
Grammatical clues	<ul style="list-style-type: none"> ▶ Make all options singular or all options plural. ▶ Use closed lead-ins.
Correct answer stands out	<ul style="list-style-type: none"> ▶ Revise options to equal length. Remove language used for teaching points and rationales.
Word repeats (clang clue)	<ul style="list-style-type: none"> ▶ Replace repeated word in either stem or option. OR ▶ Use repeated word in all options.
Convergence	<ul style="list-style-type: none"> ▶ Revise options to balance use of terms.

CHAPTER 4: ITEM ANALYSIS AND INTERPRETATION OF RESULTS

Item analysis can provide useful information about the performance of items or questions for a given group of test-takers. Item analysis traditionally involves the application of a routine set of analyses prior to calculating the final test scores and providing grades to students. This chapter covers the most common types of item analyses used in assessment, which are listed below, and provides some illustrative examples as well.

- ▶ Analysis of item difficulty
- ▶ Analysis of item discrimination
- ▶ Analysis of item options
- ▶ Comparative analysis of test-taker groups

ANALYSIS OF ITEM DIFFICULTY

Often during the item-writing process, item writers make assumptions about how easy or difficult a particular item will be based on the content area or clinical scenario used in the item. However, test-takers often confound these expectations and respond to questions in unexpected ways. Thus, the first analysis for any test item is to calculate the difficulty level of that item, using the response data. The most common classical test theory index of difficulty is the p -value, or percent-correct value. This is defined as the percent of overall test-takers who got a certain item correct. Lower p -values indicate lower percentages and more difficult items, while higher p -values indicate easier items. These values are always positive and can be represented as a percent or a proportion, so that “20” and “.20” are both acceptable ways of reporting that 20% of the test-takers got a certain item correct (once a method is chosen, it should be used consistently across all items). The observed p -value is often in and of itself meaningful, but the real value lies in comparing it to the item writer’s or test developer’s expectations. Was the item as easy or difficult as expected? Was the item unusually easy or difficult? Items that are very easy (p -value $>.95$) or very difficult (p -value $<.30$) do not provide much information about the population as a whole and may indicate that the item content is not a good match for the test-takers’ proficiency. Are unusually high or low p -values showing up on certain topics or content areas? This can result from test-takers who have completely mastered the material or not learned it at all. A high-quality assessment will contain items that, in addition to covering an appropriate range of topic areas, will represent a range of difficulties as well.

ANALYSIS OF ITEM DISCRIMINATION

A good item is one that discriminates between test-takers who know the material and those who do not. In practical terms, the index of discrimination can be computed as the correlation of test-taker performance on the item with performance on the test as a whole (where the overall test score used to estimate the correlation might include or exclude that item). Indices of item discrimination include correlation coefficients such as the biserial and point-biserial correlation; either estimate is appropriate for correlating performance on a single item, scored right-wrong, with a continuous test score. These are also known as item-total correlations. Biserial and point-biserial estimates range from -1.0 (perfect negative discrimination) to $+1.0$ (perfect positive discrimination).

Large, positive item-total correlation values indicate that the test-takers who got that item correct tended to do well on the test as a whole, so the item discriminates well. These are the most desirable types of items. When an item-total correlation is close to zero, there is little to no relationship between item performance and overall test performance, meaning that the item does not provide much additional information for rank ordering test-takers on the performance scale. When an item-total correlation is negative, this indicates that test-takers who did worse on the test overall actually had a higher chance of getting the item correct than those who did better on the test. There are several factors that can explain a zero or negative item-total correlation. The item might be measuring a different construct than the rest of the test, so that performance on that item essentially has no relation to performance on the other items. There might be an obvious flaw in the item that lower-scoring test-takers are using to guess effectively or that is causing most of the test-takers to have to guess the answer (rightly or wrongly). Finally, an item that is keyed incorrectly will have, in addition to a very low p -value, a negative correlation estimate.

ANALYSIS OF ITEM OPTIONS

An item writer should always review the performance of the incorrect options; this is known as option analysis. There are several questions to be asked during option analysis. Were any of the options not selected? This is a sign that these options were not plausible or could be ruled out due to a structural flaw or the sophistication of test-takers, and thus may need to be rewritten. Was any incorrect option chosen more often than expected, or chosen more often than the key? If the option was somewhat more likely to be chosen, this is an indication that the item could have more than one right answer; if much more likely, this is a sign that the item is probably miskeyed. Just as the keyed option should perform as expected (in the sense that the item difficulty should be in line with expectations), so should the other options. While test-takers can learn how to review and rule out incorrect options, the test developer should take notice if large numbers of test items have many distractors that are so implausible they are rarely or never chosen. If an option that is expected to be an easy exclusion or is expected to be a challenging, plausible distractor performs contrary to expectations, the item should be reviewed for structural soundness and content.

COMPARATIVE ANALYSES OF TEST-TAKER GROUPS

Comparative analyses of test-taker groups fall into two categories: (1) grouping test-takers within item by overall test performance and (2) comparing item performance across test-taker groups. The first type, within-item analysis, involves classifying students by overall test performance into a small set of groups, where sample sizes are sufficiently large for each group. A common grouping is known as High/Low, where the top 50% of the students are placed in the High group and the bottom 50% are placed in the Low group, and item difficulty and option analysis are evaluated separately by group within items. Another type of High/Low grouping compares those test-takers at the very top and bottom of the score distribution. Some item analysis research suggests that comparing the top 27% and bottom 27% provides the most useful information; in practice, this is often rounded off to the top 25% and the bottom 25%. For very large numbers of test-takers, groups can also be divided into quartiles (four groups of 25% each) or quintiles (five groups of 20% each) and each group can be compared with all the others. While item-level estimates of difficulty and discrimination are usually done on the total group, option analysis is most informative if conducted on subgroups such as High/Low.

The second type of comparative analysis, cross-group analysis, requires the grouping of students by some type of variable that would be expected to impact overall test performance; for example, in a class of first- and second-year students, the groups could be based on student year. Then, students within each year could be further grouped by performance, so that, for example, p -values and option analysis for the High groups could be compared across first- and second-year students. Another way to classify test-takers is to calculate item analysis statistics for the same items over time, using equivalent groups of test-takers. A big change in p -value or discrimination for an item over time for first-year students taking the same course in subsequent years could

indicate that the item has become “exposed” (known beforehand to test-takers), that the clinical information in the item is no longer accurate, or that the topic is no longer being taught.

Example Items and Analysis

The following are example item analysis results from five items; each illustrates a common scenario. The item text is not presented here, only the analysis. For each example, students were divided into High and Low groups based on being in the top 25% and bottom 25% of performance on the total test (where performance includes the item in question). Typically, item analysis output includes all the estimates mentioned in this chapter—comparative grouping of students, a measure of item difficulty, a measure of discrimination, and responses by option to allow for option analysis. For each of the following sample items, the percentage of test-takers in the High and Low groups selecting each option is shown. The total row shows the percentage of the total group who selected each option.

Item #1

GROUP	A	B*	C	D	E	F
HIGH	1	1	91	4	1	2
LOW	20	6	51	14	6	3
TOTAL	9	2	76	8	3	2

p -value: 2

discrimination index: -.21

Interpretation (Item #1): The asterisk on option B indicates that B was keyed as the correct answer, but only 2% of the students answered correctly, and only 1% of the High group answered correctly. This is the typical option pattern observed for an item that is miskeyed. If the answer is truly option B, the item is very difficult and the discrimination index is negative. The correct answer is almost certainly C, but a content expert should review the item for verification. If the correct answer is keyed as C, the p -value becomes 76% and the discrimination index becomes positive. These are both excellent values from a statistical perspective, and there is no reason to make any further changes before scoring the item or using it in future tests.

Item #2

GROUP	A	B	C*	D	E	F
HIGH	0	1	90	3	3	3
LOW	0	1	60	25	8	6
TOTAL	0	1	74	12	7	6

p -value: 74

discrimination index: +.33

Interpretation (Item #2): 90% of the High group and 60% of the Low group selected the correct answer, with an overall percent-correct of 74%. These are good statistics, because this item is of appropriate difficulty and does a good job of discriminating between those who know the material and those who don't. The responses to the keyed option (C) suggest the same. One conclusion of the option analysis is that A and B do not appear to be very plausible or useful distractors, so these could potentially be rewritten for future versions of the item. Keep in mind that revising options to make them more plausible can change the difficulty and discrimination of the item, sometimes in unpredictable ways.

Item #3

GROUP	A	B	C*	D	E	F
HIGH	44	1	50	2	1	2
LOW	20	15	21	22	20	2
TOTAL	32	7	34	14	11	2

p-value: .34

discrimination index: +.30

Interpretation (Item #3): 50% of the High group and 21% of the Low group selected the correct answer. This is a difficult item, so a content expert should review it to ensure that the key is correct—there is the potential for option A to be a second correct answer. If the item was not intended to be this difficult, it is important to review the structure of the item, as there may be flaws that make the item confusing for the test-taker. However, if the item was intended to be this difficult, the content expert agrees that the keyed option is the single correct answer and A is definitely an incorrect answer, then the item can be scored as is.

Item #4

GROUP	A	B	C*	D	E	F
HIGH	18	10	51	17	2	2
LOW	24	24	21	25	4	2
TOTAL	22	17	34	22	3	2

p-value: .34

discrimination index: +.30

Interpretation (Item #4): The High/Low group breakdown on option C is identical to Item #3, but this item is less likely to have potential problems. In contrast to the previous sample item, those test-takers who don't know the correct answer are more evenly spread across the other distractors, which is often a sign that they don't know the answer and have to guess among all the options. For the three distractors, A, B, and D, more test-takers in the Low group chose the distractor than test-takers in the High group. Of course, if the item was not intended to be difficult, it would still be desirable to review options A, B, and D for correctness and clarity.

Item #5

GROUP	A	B	C	D*	E
HIGH	10	43	5	40	2
LOW	23	36	12	26	3
TOTAL	17	43	7	31	2

p-value: .31

discrimination index: -.09

Interpretation (Item #5): The *p*-value is low and the discrimination is negative, indicating that there is a problem with the item. Both the High and Low groups are more likely to select option B than the option keyed as correct, which is D. This is a classic example of an item that likely has two correct answers. This item should be reviewed by a content expert and should not be scored until it is reviewed, since something about the item stem or options is convincing even the High performers that the key is an answer other than D.

SECTION 2:

WRITING
ONE-BEST-ANSWER
ITEMS FOR THE
FOUNDATIONAL
(BASIC) AND
CLINICAL
SCIENCES



SM

CHAPTER 5: BASIC RULES FOR WRITING ONE-BEST-ANSWER ITEMS

RULE 1: Each item should focus on an important concept or testing point.

As a health care provider or educator involved in the development of an examination, you may be asked to write items to assess test-taker knowledge of a particular domain. What do you want the test-taker to know or demonstrate? The topic of the item usually results from the examination blueprint, which is the outline of the major topics to be covered. For instance, if an examination is intended to assess knowledge of the cardiovascular system, the blueprint might have two dimensions: 1) disease-based (eg, hypertension, ischemic heart disease, systolic heart failure), and 2) task-based (eg, assessment of foundational science principles, diagnosis, history, prognosis). The blueprint would likely include topics along both dimensions and might call for six items on hypertension, four on systolic heart failure, two on diastolic heart failure, ten on ischemic heart disease, and so on. Along the task dimension there might be a similar distribution of topics. A clear and comprehensive blueprint or other set of test specifications should always be available so that item writers can stay focused on the important topics and write a sufficient number of items for each topic.

RULE 2: Each item should assess application of knowledge, not recall of an isolated fact.

The first step in writing an item is to develop an appropriate stimulus to introduce the topic, such as a clinical or experimental vignette, to provide context to the question being asked. If there is no such stimulus, the resulting item will generally be assessing knowledge recall. Recall items make it difficult for the educator to assess any higher level within Bloom's taxonomy, such as "application of knowledge." For instance, an item consisting of one sentence, "Which of the following medications is used to decrease preload in systolic heart failure?" would assess only the recall on the mechanisms of action of a list of pharmacotherapeutic agents.

It can be helpful to use actual patient scenarios that you previously encountered as a source of ideas for items and vignettes. However, you should avoid relying on or adhering too closely to patient cases because these often have atypical features that may divert from a typical or representative case and lead to confusion. Additionally, in some instances, such as the example with systolic heart failure, there will be an additional step that you must keep in mind: you should consider the underlying cause of the heart failure. Patient demographics, past medical history, and other factors will differ depending on the cause of the condition. Patients with systolic heart failure from a viral cardiomyopathy versus from ischemic heart disease may have different demographics and a different history (eg, a younger patient with a viral illness preceding the onset of heart failure symptoms as compared to an older patient with risk factors for ischemic heart disease).

The details of the vignette should be guided by the level of the test-taker. Here are two examples for test-takers with two levels of education/experience:

Test-taker with Less Education/New Experience: A systolic heart failure vignette for a second-semester first-year medical student would include very typical features and classic symptoms: shortness of breath with physical activity that improves with rest; awakening at night short of breath, relieved by sitting up; pedal edema; and pertinent negatives such as the absence of chest pain. Risk factors might include an upper respiratory illness two weeks ago or a history of heavy alcohol ingestion over 20 years.

Test-taker with More Advanced Education/Skills: A test-taker, such as one sitting for a specialty certifying examination, would be able to work through a vignette that included some atypical features, as is the case with many actual patients. The demographic information may or may not be significant for the more advanced test-takers. For instance, every patient lives somewhere, and many will have a current or past occupation that may or may not be related to the cause of their illness. In a vignette for a 30-year-old man with shortness of breath and wheezing in which the diagnosis is asthma, the demographic information might or might not be related to the diagnosis. The patient might be a farmer, but the most likely diagnosis is still asthma and not farmer's lung or silo-filler's lung.

RULE 3: The item lead-in should be focused, closed, and clear; the test-taker should be able to answer the item based on the vignette and lead-in alone.

The next step in item writing is to phrase the question with the use of a lead-in, where the accompanying vignette allows the lead-in to be focused on the patient, such as, “Which of the following is the most appropriate next step in management?” or “Which of the following is the most likely diagnosis?” An open-ended lead-in such as, “The diagnosis in the patient is:” should be avoided. The lead-in should be a single, closed, clear question. Ideally, after reading the vignette and the lead-in, a test-taker should be able to answer the item without seeing the options. Another reason to use a closed lead-in is because it helps to avoid certain item flaws, such as grammatical cueing.

RULE 4: All options should be homogeneous and plausible to avoid cueing to the correct option.

Homogeneity:

At this point in item writing, the patient-based, closed lead-in created with Rule 3 in mind will direct the focus and grammatical form of the answer options. Maintaining a consistent focus and parallel format among the answer options results in homogeneity, which allows test-takers to weigh each option within a single mindset without construct-irrelevant distractions. For example, in response to “Which of the following is the most likely cause of this patient’s condition?” a list of answer options in which all choices are diagnoses (eg, tuberculosis, meningitis, etc.), is easier to process than a list containing both diagnoses and underlying pathogens (eg, tuberculosis, *Neisseria meningitidis*, etc.).

Plausibility:

The correct answer should always be the “most” correct of the answer options, but the distractors should be plausible enough to entice test-takers who do not know the correct answer. Otherwise, test-takers can arrive at the correct answer by eliminating distractors based on their improbability within the context of the patient scenario.

When writing answer options, start by generating the correct answer for the lead-in. Generating parallel and plausible yet incorrect distractors is more challenging. For questions regarding diagnosis, the topic area may be the answer—if you are assigned to write two items on community-acquired pneumonia (CAP), one item on

CHAPTER 6: TESTING APPLICATION OF FOUNDATIONAL (BASIC) AND CLINICAL KNOWLEDGE

CHOOSING THE TOPICS TO TEST

The content of an exam should be driven by the purpose of that exam and the test-taker population. Who is being tested and how will the scores be used? For example, the USMLE system is designed for use by state medical licensing authorities in their decision to grant a general licensure for allopathic and international graduate physicians or other providers in the United States. The focus is to assess knowledge of content that is necessary for the practice of medicine by the undifferentiated physician or other provider; items might be included on USMLE that assess knowledge not uniformly taught in medical school. Conversely, topics within some medical schools' curriculum might be omitted from the exam. The analogy for individual schools and courses within schools is to determine the student test-taker population and purpose of the scores. If the purpose of the content is to test for specialty or subspecialty certification for physicians or other providers, the content and its inferences should be geared toward the required minimum competence in that specialty upon beginning practice. An exam that is intended for formative feedback at a midpoint of a course will have a different focus and different content from an exam to determine end-of-clerkship grades.

DETERMINING LEVEL OF COGNITION TO ASSESS

Items can be grouped into two general categories, based on the cognitive task required of the test-taker:

1. **Recall of a Fact:** An item that assesses rote memory of a fact (without requiring its application).
2. **Application of Knowledge:** An item that requires a test-taker to apply knowledge to reach a conclusion, make a prediction, or select a course of action that does not depend on memory alone.

Items that test recall of a fact require test-takers to read an item and to recall isolated facts, concepts, and principles or to recognize previously encountered situations (eg, experiments, patient encounters, case studies). These items often begin by citing a disease and then asking what patient findings are expected. For example, "Which of the following findings is most likely to be seen in postsurgical patients with pulmonary embolism?" is an item structured similarly to most textbook questions. The test-taker could look up the disease and find the answer in a single paragraph. From a practical standpoint, these items also seem clinically backwards—patients would not tell their provider what disease they have and then ask the provider to determine the signs and symptoms.

Application of knowledge items, on the other hand, require test-takers to read an item and identify relevant information, interpret that information in a certain context, integrate that information with what they already know, and then answer the question posed. Vignette-based items (items that include a detailed patient or experimental scenario) often provide a vehicle for eliciting the demonstration of these higher-order thinking skills. Some examples of these application of knowledge items can be found throughout this book.

Determining the cognitive task for an item – recall vs application of knowledge – depends on the intended end-use of the item. The use of recall items may be best for formative assessment purposes or the evaluation of simpler concepts that might not lend themselves to clinical or experimental scenarios. For a medium-to-high-stakes summative examination, use of vignette-based items that require higher-order thinking skills and application of knowledge would be preferable to simple recall items.

Figure 2. Example of a clinical scenario as a recall item vs an application of knowledge item

A woman is diagnosed with venous thromboembolism. Which of the following is the most appropriate treatment?

- ▶ This is an example of a recall of a fact item, which requires examinees to simply recall the treatment of venous thromboembolism.

USE A VIGNETTE AS A VEHICLE TO TEST APPLICATION OF KNOWLEDGE

A 47-year-old woman comes to the emergency department because of shortness of breath and left lower extremity pain with ambulation. Yesterday she returned from Europe after a 10-hour flight. She has no remarkable medical history and takes only an oral contraceptive daily. Vital signs are within normal limits. Physical examination shows asymmetrical swelling of the left calf of greater than 2 cm compared to the right side. The remainder of the physical examination discloses no abnormalities.

- ▶ This vignette provides more clinical context, requiring the examinee to recognize the historical and physical presentation of venous thromboembolism.
- ▶ From this, a variety of questions could be asked that require the examinee to interpret data, including the following:
 - ▷ Diagnosis
 - ▷ Next step in determining the diagnosis (which requires clinical suspicion of diagnosis and the next test to order)
 - ▷ Treatment (which requires clinical suspicion of diagnosis and determination of the most appropriate next step in management)
 - ▷ Mechanism (which requires clinical suspicion of diagnosis, the most appropriate treatment, and the mechanism of that treatment)

Benefits of Application of Knowledge Item Type

Items with a clinical vignette to assess application of knowledge have several benefits:

1. The authenticity of the examination is greatly enhanced by using items that require test-takers to integrate information to “solve” clinical problems.
2. The items are more likely to focus on important information, rather than trivia.
3. These items help to identify those test-takers who have memorized a substantial body of factual information but are unable to use that information effectively in clinical situations. Test-takers need to differentiate relevant from irrelevant information in the item.

The Shape of a Good Item

A well-constructed one-best-answer item will have a particular silhouette as shown in the illustration below. A rich clinical scenario serves as the stem, and all of the options are listed in a concise and uniform manner. The stem should include all relevant facts; no additional information should be provided in the options.

Tell your story here in the **vignette**.

Pose your question here in the **lead-in**.

- A.
- B. Insert your answer **option set**
- C. here, making sure it follows
- D. the **“cover-the-options”** rule.
- E.

Make sure the item stem adheres to the following rules:

- ▶ Focuses on important concepts rather than trivial facts
- ▶ Can be answered without looking at the options
- ▶ Includes all relevant facts; no additional data should be provided in the options
- ▶ Is not “tricky” or overly complex
- ▶ Is not negatively phrased (eg, avoid using “except” or “not” in the lead-in)

NOTES

THE PATIENT CHART/TABLE FORMAT

A variation of the clinical vignette uses a “patient chart” format or table format for providing details of a patient case/scenario. These are best used when a lot of detail is provided about the patient and/or the case. However, use caution when constructing complex scenarios (eg, trauma cases, patients with multiple comorbid conditions or hospital admissions) because timing and condition changes could be lost in this format.

Example of Patient Chart Format



PATIENT INFORMATION

Age: 62 years
Gender: M, self-identified
Ethnicity: unspecified
Site of Care: office

HISTORY

Reason for Visit/Chief Concern: “My legs hurt when I walk and it’s getting worse.”

History of Present Illness:

- ▶ 3-month history of worsening leg pain
- ▶ pain exacerbated by walking; peak intensity after one block
- ▶ pain resolves completely with rest
- ▶ pain rated 4/10 at worst

Medications:

- ▶ lisinopril
- ▶ metoprolol
- ▶ furosemide
- ▶ glyburide
- ▶ lovastatin

Past Medical History:

- ▶ hypertension
- ▶ mild angina
- ▶ type 2 diabetes mellitus

Allergies:

- ▶ no known drug allergies

Psychosocial History:

- ▶ has smoked one-half pack of cigarettes daily for 44 years

PHYSICAL EXAMINATION

Temp	Pulse	Resp	BP	O ₂ Sat	Ht	Wt	BMI
36.0°C (96.8°F)	72/min	14/min	140/90 mm Hg	-	164 cm (5 ft 5 in)	90 kg (198 lb)	33 kg/m ²

EXAMPLE CONTINUES ON FOLLOWING PAGE



- ▶ **Appearance:** no acute distress
- ▶ **HEENT:** funduscopy shows grade 2/4 arteriovenous nicking
- ▶ **Neck:** no jugular venous distention
- ▶ **Pulmonary:** clear to auscultation; mildly diminished lung sounds
- ▶ **Cardiac:** no bruits; distant heart sounds
- ▶ **Abdominal:** obese; no tenderness, guarding, masses, bruits, or hepatosplenomegaly
- ▶ **Extremities:** no joint erythema, edema, or warmth; no hair on toes; no femoral bruits; dorsalis pedis, radial, and femoral pulses intact
- ▶ **Neurologic:** sensation to vibration intact

Question: Which of the following is the most appropriate diagnostic study?

- A. Ankle brachial index*
- B. Arteriography
- C. ECG
- D. Echocardiography
- E. MUGA scan

THE F-TYPE FORMAT

In addition to A-types, clinical vignettes work very well with F-types, also known as case-based or sequential item sets. F-types usually include two to three items associated with a clinical scenario that unfolds over time. As the case advances, additional information is provided that requires examinees to decide the next step or make a decision at that point in time.

Rules for Writing F-type Sets

1. Include enough rich clinical context in the initial scenario to support unfolding of the case in subsequent items.
2. Advance the clinical situation in time and provide additional patient information to bring the test-taker back to baseline, if necessary.
3. Imply a response to the previous item but make sure not to turn the next item into a recall type.



Example of F-type Set

INITIAL SCENARIO

A 35-year-old woman is brought to the emergency department because of worsening pain and swelling of the right knee for the past 2 days. She has been taking acetaminophen for the knee pain during the past 2 days, but the pain is worse today. She has not had any trauma to the knee or any previous problems with her joints. She is otherwise healthy, and she currently takes an oral contraceptive. She is sexually active and has a 10-year-old son who lives with her. She is 160 cm (5 ft 3 in) tall and weighs 52 kg (115 lb); BMI is 20 kg/m². Temperature is 37.9°C (100.2°F). The right knee is erythematous, swollen, and tender, and there is pain with movement. No other joints are affected. X-ray of the knee shows an effusion but no structural abnormalities of the joint.

ITEM 1

Which of the following is the most appropriate next step in diagnosis?

- A. Arthrocentesis of the knee*
- B. Blood cultures
- C. Complete blood count
- D. MRI of the knee
- E. Urine cultures

EVOLVING SCENARIO 2

Arthrocentesis is done. The synovial fluid is cloudy. Gram stain is negative. Analysis of the synovial fluid shows a leukocyte count of 120,000/mm³ and 90% neutrophils.

ITEM 2

Which of the following is the most appropriate additional test on the synovial fluid?

- A. Culture for bacteria*
- B. Glucose concentration
- C. Polarized light microscopy
- D. Protein concentration

NOTES

GUIDELINE(S) FOR WRITING ITEM LEAD-INS (see Appendix B for more detail)

The lead-in should consist of a single, clearly formulated question so that the test-taker can answer without looking at the options. As mentioned previously, satisfying the “cover-the-options” rule is an essential component of a good question. The following stem example provides sufficient information and can be answered without referring to the options:

A 52-year-old man has had increasing dyspnea and cough productive of purulent sputum for 2 days. He has smoked one pack of cigarettes daily for 30 years. Temperature is 37.2°C (99.0°F). Breath sounds are distant with a few rhonchi and wheezes. Leukocyte count is 9000/mm³ with a normal differential. Gram stain of sputum shows numerous neutrophils and gram-negative diplococci. X-ray of the chest show hyperinflation. Which of the following is the most likely diagnosis?

Item Example With and Without Vignette

The following trio of items were administered on USMLE in the past and performed quite differently across the various formats, especially for low-performing test-takers. The grid beside each item shows the percentage of High (top 20%) and Low (bottom 20%) students who selected each option. Almost all of the High group (99%) and the Low group (90%) selected the correct option (indicated by the asterisk) in the non-vignette format. The short- and long-vignette formats were not markedly more difficult for the High group but were for the Low group; the correct answer was selected by 82% of the Low group in the short-vignette format and 66% in the long-vignette format. See Chapter 4 for a more in-depth discussion of item analysis.

NO VIGNETTE

The most likely renal abnormality in children with nephrotic syndrome and normal renal function is

- A. acute poststreptococcal glomerulonephritis
- B. hemolytic-uremic syndrome
- C. minimal change nephrotic syndrome*
- D. nephrotic syndrome due to focal and segmental glomerulosclerosis
- E. Schönlein-Henoch purpura with nephritis

	A	B	C*	D	E
HIGH	1	0	99	0	0
LOW	8	1	90	1	0

SHORT VIGNETTE

A 2-year-old child has a 1-week history of edema. His blood pressure is 100/60 mm Hg, and there is generalized edema and ascites. Serum concentrations are: creatinine 0.4 mg/dL, albumin 1.4 g/dL, and cholesterol 569 mg/dL. Urinalysis shows 4+ protein and no blood. Which of the following is the most likely diagnosis?

(Same option set as above)

	A	B	C*	D	E
HIGH	0	0	98	2	0
LOW	6	2	82	9	1

LONG VIGNETTE

A 2-year-old child developed swelling of his eyes and ankles over the past week. Blood pressure is 100/60 mm Hg, pulse 110/min, and respirations 28/min. In addition to swelling of his eyes and 2+ pitting edema of his ankles, he has abdominal distention with a positive fluid wave. Serum concentrations are: creatinine 0.4 mg/dL, albumin 1.4 g/dL, and cholesterol 569 mg/dL. Urinalysis shows 4+ protein and no blood. Which of the following is the most likely diagnosis?

(Same option set as previous page)

	A	B	C*	D	E
HIGH	0	1	98	1	0
LOW	10	9	66	10	5

Although the third item listed is labeled “long vignette,” it is still relatively short in length. Clinical knowledge and science exams require test-takers to demonstrate proficiency in sorting through patient information, synthesizing the important findings, and reaching a conclusion. As a result, these items may have extraneous information as well as the essential information to answer the question. If there is concern about the vignette length, it is possible to synthesize findings with a statement such as, “The family history is noncontributory.”

Additional Points on Vignettes

Verbosity, Window Dressing, and Red Herrings

Many educators stress the importance of writing items that are as short as possible as a method of avoiding verbosity, “window dressing” (extraneous material not needed to answer the item), and “red herrings” (information designed to mislead the test-taker). However, it is possible to avoid these traps while writing good-quality clinical vignettes that stress application of knowledge by asking test-takers to make clinical decisions rather than simply recall isolated facts. These items are designed to reflect “real-life” tasks by challenging test-takers to first identify the findings that are important and then integrate those findings into a diagnosis or clinical action. These items often require multiple steps in the cognitive process. The vignettes tend to follow a standard structure and pose questions that are clinically natural, and the use of a template allows for development of high-quality vignettes with a lower risk of adding too much verbiage or unnecessary or confusing information.

Use of Real Patients

As mentioned previously, item writers should be careful when basing vignettes on real patients, particularly for tests aimed at introductory learners or test-takers. As a general rule, real patients are complicated, and the elements that are complicated are not always those that are important for assessment. As noted earlier, it is fine to sometimes include window dressing, such as incidental findings, but item writers should note that real patients often have red herrings among their findings.

Patients Who Convey an Unreliable History

Ideally, patients in vignettes should tell the truth, or the provider’s interpretation of the patient’s story should be provided. Providers use multiple cues to determine how truthful a patient is, and many of these cues cannot be translated into written form. Thus, an item may describe a patient’s alcohol consumption as, “The patient drinks 16 ounces of beer with dinner each night” or, “The patient’s description of his alcohol consumption is contradictory.” Do not write something that requires an interpretation of veracity, such as, “The patient ‘claims’ to drink only one bottle of beer each night.”

Patient Characteristics in Item Creation

Characteristics of a patient such as age, sex, gender identity, disability, socioeconomic status, native language, country of origin, and/or occupation are sometimes mentioned within case vignettes in test items. Some patient characteristics (PC) may be important inputs into the diagnostic reasoning process. Others may lead to incorrect conclusions and misdiagnoses. Among the latter are characteristics that could potentially be associated with harmful patient stereotypes.

When creating items, be mindful of the notion that race is a social construct not linked to biology or susceptibility to disease. This is similarly true of ethnicity and culture, heritage, or even country of origin. Ancestry, if known, may be biologically important, and thus may be relevant to factors relating to health and disease. In addition, when and if these characteristics are to be considered for inclusion in your items, they should be considered based on patient self-report, not the assumption of a health care provider.

PC can be described and included in vignettes if they:

- ▶ are clinically relevant and/or could aid in distractor quality.
- ▶ are necessary for the examinee to better understand the context in which the patient is being seen (the item would be unreasonably difficult if excluded).
- ▶ add to the overall exam-level representativeness of the referenced patient population.
- ▶ increase the probability of detection, diagnosis, or recognition of an otherwise rare condition.
- ▶ do not contain negative stereotypes.

Test items should be carefully designed to measure meaningful and plausible testing points (eg, diagnosis, management, etc.), without the influence of assumptions, bias, or stereotypes. When examinees select the correct (keyed) response, they are given credit because they are demonstrating what the examination item is designed to measure. Health professionals and educators creating assessments should follow guidelines that encourage thoughtful consideration of PC, while at the same time strive to promote diversity and present patients who reflect the population served by your examinees.

Examples of Challenging Patient Characteristics Scenarios

EXAMPLE A:

A 2-year-old girl is brought to the office because of bilateral hand swelling and splenomegaly..... Her family identifies as African American.

In this example, the lack of inclusion of any patient characteristics (PC) could make clinical reasoning very difficult. However, given sickle cell disease's prevalence in Latin America, the Middle East, and Southern European regions such as Turkey, Greece, and Italy, it would be reasonable to change the PC to one that represents one of the regions where this disease association exists, but is less well known.

“... SOME PATIENT CHARACTERISTICS (PC) MAY BE IMPORTANT INPUTS INTO THE DIAGNOSTIC REASONING PROCESS. OTHERS MAY LEAD TO INCORRECT CONCLUSIONS AND MISDIAGNOSES...”

EXAMPLE B:

A 25-year-old man with known sickle cell disease comes to the emergency department because of 2 days of severe back pain. His medical record suggests he was seen in the emergency department 5 days ago and given a prescription for opioid analgesia of 3 days' supply..... The patient identifies as African American.

In contrast to example A, in which PC for a patient with sickle cell disease is relevant to the item, the PC in example B is not needed and may perpetuate harmful stereotypes. Such patient characteristics should not be included when writing an item about drug seeking behavior. Furthermore, adding race/ethnicity is not likely needed when the diagnosis is already provided in the item.

EXAMPLE C:

A 23-year-old woman comes to the office because of bloating and abdominal pain after ingesting dairy products. She states that she is of Native American descent.

Nearly 100% of Native American people have lactose intolerance in adulthood. The condition also occurs in other populations with significant frequency (25% of European people and 50% to 80% of people of Hispanic descent). Here, the patient characteristics could be changed from Native American to another ethnicity, but one could argue that the PC are not needed for clinical reasoning given the patient's symptoms after ingesting dairy. However, PC may still be included (eg, Native American) if trying to add to the overall exam-level representativeness of the reference patient population.

EXAMPLE D:

A 46-year-old woman with cough, dyspnea, and chest pain with bilateral hilar adenopathy comes to the office..... She identifies as African American.

Even though sarcoidosis is four times more common in people of African ancestry, its overall rate is still only 10/10,000. Therefore, the PC may not be necessary here—or the item as written should be omitted if it does not contain sufficient additional information (beyond PC) for clinical reasoning.

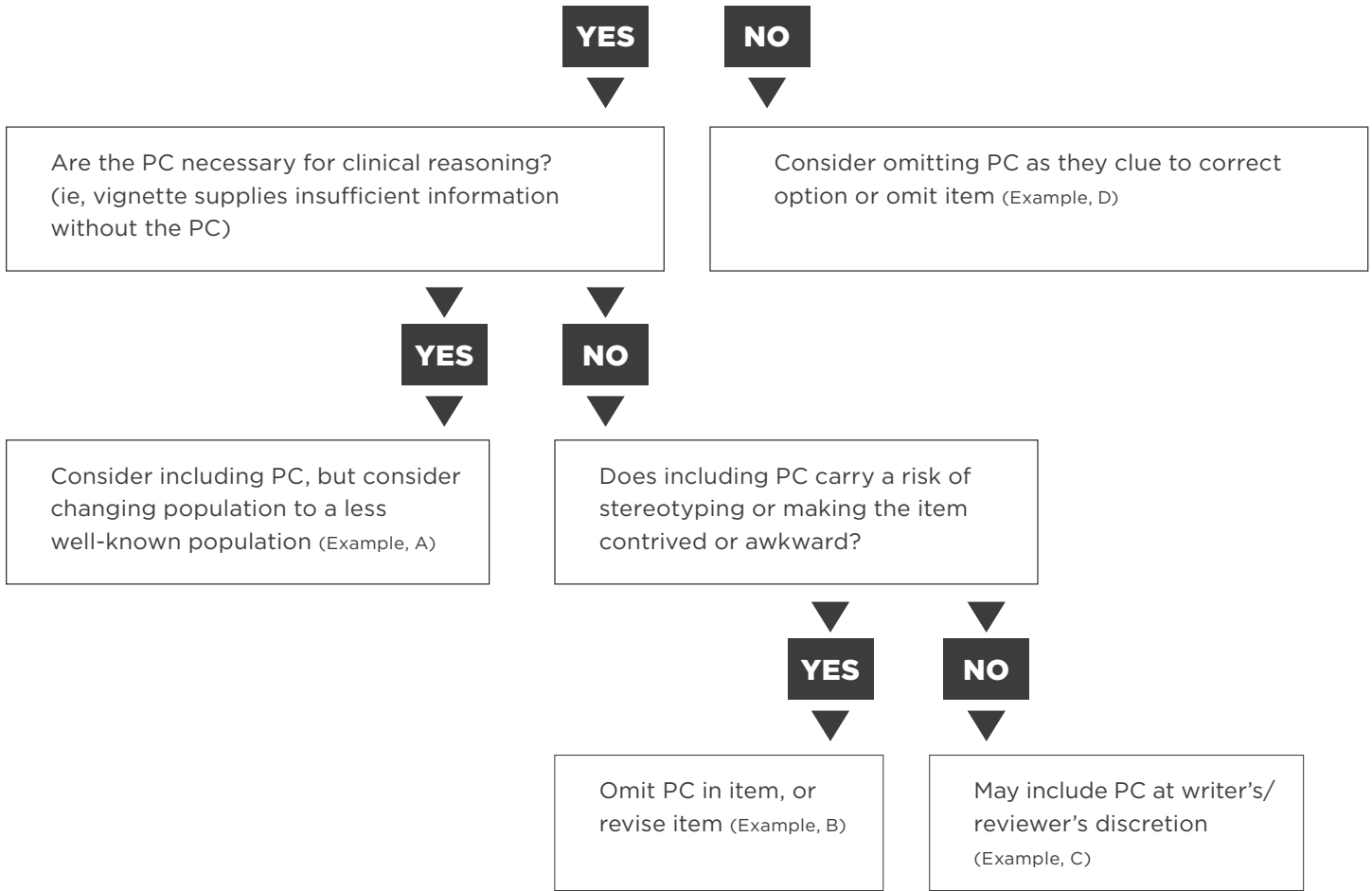
EXAMPLE E:

A 22-year-old man who is from Greece comes to the office with hypochromic, microcytic anemia...

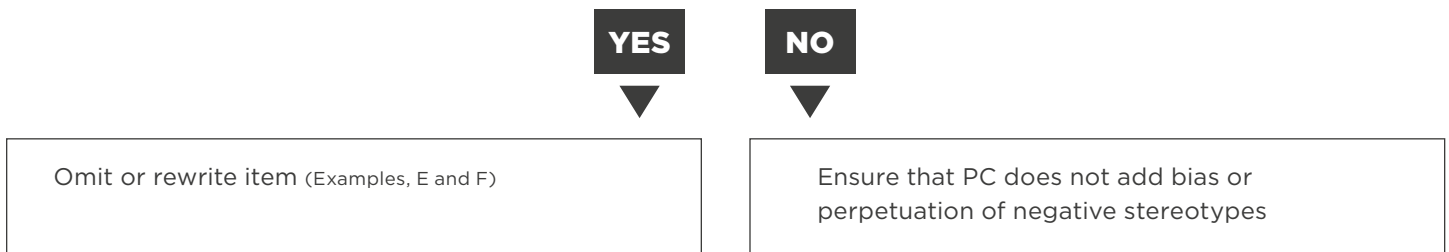
Thalassemia is most prevalent in populations with Mediterranean ancestry, to the point that the disease's name is derived from the Greek language. Providing this particular PC would make the item too easy.

GUIDELINES FOR ITEM WRITERS ON THE USE OF PATIENT CHARACTERISTICS (PC)

Does the disease or condition in the item occur in a given population with a reasonable degree of frequency?



**Does inclusion of the patient characteristics make the item too easy?
Alternatively: Does omission of the patient characteristics make the item too difficult?**



STRUCTURING ITEMS TO FIT TASK COMPETENCIES

A set of defined task competencies will assist the item writer in focusing his or her intended testing point. Each competency requires a slightly different approach to item writing. Some sample lead-ins and example items to guide item-writing efforts for each physician (or other provider) task competency are provided below. Additional lead-ins can be found in Appendix B: Sample Lead-ins Based on Provider Task Competencies.

Foundational (Basic) Science

Foundational science comprises items that require understanding and application of basic science. These items should require clinical knowledge as well as knowledge of one or more foundational science principles that would likely have been learned during preclinical study and reinforced during clinical rotations. Mechanisms of disease is an example of a competency within the foundational science category. Items in this competency should evaluate test-takers' knowledge of pathophysiology in its broadest sense, including etiology, pathogenesis, natural history, clinical course, associated findings, complications, severity of illness, and intended or unintended effects of therapeutic interventions. These items should be framed in a clinical context. In general, the writer should open items on mechanisms of disease with a clinical vignette featuring a patient with specified symptoms, signs, history, and lab study findings. The following lead-ins are examples of those used to test foundational science principles:

- ▶ Which of the following is the most likely cause/mechanism of this effect?
- ▶ Which of the following is the most likely infectious agent?
- ▶ Which of the following is the most likely explanation for these findings?
- ▶ Which of the following is the most likely location of this patient's lesion?
- ▶ Which of the following is the most likely pathogen? (Interpretation of foundational-science-based information, such as Gram stain results, should be required in the vignette to differentiate the competency being tested from that of a clinically based most likely diagnosis item.)
- ▶ Which of the following findings is most likely to be increased/decreased in this patient?
- ▶ A biopsy specimen is most likely to show which of the following?
- ▶ This patient most likely has a defect in which of the following?
- ▶ This patient most likely has a deficiency in which of the following enzymes?
- ▶ Which of the following cytokines is the most likely cause of this condition?
- ▶ Which of the following structures is at greatest risk for damage during this procedure?
- ▶ The most appropriate medication for this patient will have which of the following mechanisms of action?

A 10-year-old girl develops gross hematuria 14 days after a sore throat. She has a blood pressure of 170/100 mm Hg and 2+ pedal and pretibial edema. Serum urea nitrogen concentration is 3.2 mg/dL. Which of the following is the most likely cause?

- A. Acute postinfectious glomerulonephritis*
- B. Microscopic polyangiitis
- C. Minimal change disease
- D. Thin basement membrane nephropathy





A 32-year-old man has a purulent urethral discharge. A culture grows *Neisseria gonorrhoeae* sensitive to penicillin. One week after cessation of penicillin therapy, the patient has a recurrence of the urethral discharge. A culture again shows *N. gonorrhoeae* sensitive to penicillin. Both the patient and his sexual partner do not have HIV infection. Examination of the patient's sexual partner shows an anal fissure; urethral culture grows no pathogens. Which of the following is the most likely cause of the recurrence of urethral infection in this patient?

- A. Concurrent herpesvirus infection
- B. Emergence of bacterial resistance
- C. Inadequate treatment with penicillin
- D. Reinfection from sexual partner*

Diagnosis

The diagnosis competency is subcategorized into more detailed concepts: Obtaining and Predicting History and Physical Examination, Selecting and Interpreting Diagnostic Studies, Formulating the Diagnosis, and Determining Prognosis/Outcome. Sample lead-ins for the various subcategories are shown.

Obtaining and Predicting History and Physical Examination

- ▶ Which of the following factors in this patient's history most increased her risk for developing this condition?
- ▶ Which of the following additional information regarding this patient's history is most appropriate to obtain at this time?
- ▶ Which of the following is the most appropriate focus of the physical examination at this time?

Selecting and Interpreting Diagnostic Studies

- ▶ Which of the following is the most appropriate diagnostic study to obtain at this time?
- ▶ Which of the following laboratory studies is most likely to confirm the diagnosis?
- ▶ Which of the following is the most likely explanation for these laboratory findings?
- ▶ Arterial blood gas analysis is most likely to show which of the following sets of findings?

Formulating the Diagnosis

- ▶ Which of the following is the most likely diagnosis?
- ▶ Which of the following is the most likely working diagnosis?

Determining Prognosis/Outcome

- ▶ Based on these findings, this patient is most likely to develop which of the following?
- ▶ Which of the following is the most likely complication of this patient's current condition?



A 28-year-old woman comes to the clinic with palpitations that occur approximately once a week, last 1 to 5 minutes, and consist of rapid, regular heart pounding. The episodes start and stop suddenly and have not been associated with chest discomfort or dyspnea. There is no history of heart problems. She drinks two to three cups of regular coffee daily. She rarely drinks alcohol and does not smoke cigarettes. Her blood pressure is 120/88 mm Hg, and pulse is 96/min and regular. A stare and lid lag are noted. The thyroid gland is firm and 1.5 times larger than normal. There is a midsystolic click at the apex and a grade 2/6 early systolic murmur at the upper left sternal border. An ECG reveals sinus tachycardia. Which of the following is the most appropriate next step in diagnosis?

- A. Ambulatory ECG monitoring
- B. Echocardiography
- C. MUGA scan
- D. Serum TSH concentration*
- E. Urine catecholamine concentration

Management

The management competency contains a range of concepts, such as Health Maintenance and Disease Prevention, Pharmacotherapy, and Clinical Interventions/Treatments. In most items that focus on the testing point of management, the examinee must infer the patient's diagnosis to determine the appropriate management.

Health Maintenance and Disease Prevention

Items in this topic area assess the ability to evaluate risk factors, understand epidemiologic data, and apply preventive measures. Health Maintenance and Disease Prevention items commonly fall into one of the following categories: 1) screening tests, 2) constructive interference, 3) vaccines/travel medicine, or 4) emergency intervention. In general, the writer should open the items with a clinical vignette that describes a patient. In addition to physical examination findings, these vignettes may include information about vaccination history, risk factors, and family history. Information about the community may be relevant and therefore included, but the question should focus on the individual patient. Questions should NOT focus on the direct assessment of isolated facts. For example, avoid asking about the leading cause of death in some subpopulation; instead, focus on the application of this knowledge. In asking about vaccines or screening tests, consider focusing the testing point on essential concepts of vaccination or screening that do not require memorization of conflicting recommendations. The following lead-ins are examples of those used in this category:

- ▶ Which of the following vaccines is most appropriate to administer at this time?
- ▶ Which of the following is the most appropriate screening test?
- ▶ Which of the following tests would have predicted these findings?
- ▶ Which of the following is the most appropriate intervention?
- ▶ For which of the following conditions is this patient at greatest risk?
- ▶ Which of the following is most likely to have prevented this condition?
- ▶ Which of the following is the most appropriate next step in management to prevent [morbidity/mortality/disability]?
- ▶ Which of the following is the most appropriate recommendation to prevent disability from this patient's injury/condition?

- ▶ Early treatment with which of the following is most likely to have prevented this patient's condition? (It is best for options to be treatments that would not have been expected to have been done previously to avoid confounding the testing point with potential medical malpractice.)
- ▶ Supplementation with which of the following is most likely to have prevented this patient's condition?

A 15-year-old boy has had two episodes of severe anaphylactic shock following bee stings in the past. He says he sometimes forgets to carry his epinephrine injector with him. Medical history otherwise is unremarkable and he takes no other medications. Vital signs are within normal limits. Physical examination shows no abnormalities. Which of the following is the most appropriate recommendation at this time?



- A. Corticosteroid therapy during the summer
- B. Desensitization with bee venom extract*
- C. Long-term prophylactic antihistamine therapy
- D. Protective clothing
- E. Restriction to home during the summer

A 33-year-old woman, gravida 1, para 1, spontaneously delivers a 2460-g (5-lb 7-oz) female newborn at 38 weeks' gestation. The newborn has hepatosplenomegaly, patent ductus arteriosus, and cataracts. At 8 weeks' gestation, the mother developed a maculopapular rash, enlarged cervical lymph nodes, sore throat, and arthralgias that spontaneously resolved in 1 week. The subsequent prenatal course was uncomplicated. She did not receive prenatal care. Which of the following tests during pregnancy is most likely to have predicted the findings in the fetus?



- A. Amniocentesis to determine karyotype
- B. Culture for herpes simplex virus
- C. Serial rubella titers*
- D. Urinalysis for cytomegalovirus
- E. VDRL test

A healthy, moderately active 75-year-old woman is found on routine screening to have a total serum cholesterol concentration of 208 mg/dL and serum HDL-cholesterol concentration of 70 mg/dL. ECG shows no abnormalities. Which of the following dietary recommendations is most appropriate?



- A. Decreased intake of cholesterol
- B. Decreased intake of saturated fat
- C. Decreased intake of simple carbohydrates
- D. Increased intake of fiber
- E. No change in diet*

CHAPTER 7: USING MEDIA AS PART OF CLINICAL VIGNETTES

The computer-based administration of a multiple-choice exam makes it straightforward to add media to test items. There are many advantages to adding media, most notably the opportunity that images, videos, and other media provide for adding authenticity to the assessment of knowledge and skills. While text-based vignettes are well-suited to the assessment of the health sciences, the addition of media can improve an item that describes the appearance of a patient or a physical exam. In addition, the presence of media allows the item writer to assess skills that purely text-based items cannot measure well (many noncognitive skills may fall into this area). Finally, a long clinical vignette that fully describes a patient's condition may be challenging to write without including textual cues that benefit the savvy test-taker. Using media in the place of this text not only provides authenticity but also avoids giving the answer away in the description.

When writing test items that use media, the goal should be to select the media that best simulates what happens in practice. There are many media types for the item writer to consider. In order to determine whether a media type is a good simulation for what happens in practice, it is important to consider the following:

- ▶ The content area being covered or the skills being assessed: For example, if the topic area is about findings on cardiac auscultation, the image of an ECG is a natural fit for assessing that skill.
- ▶ The novelty of the media: Very novel media may require a learning curve or additional tutorial information to orient test-takers, so simplicity in accessing media could be a desirable factor.
- ▶ The memorability of the media: Media may be more easily remembered by students, which can be problematic if a limited sample of different images or videos is used for multiple classes or exams. Ideally, students would not be able to easily memorize notable features of the test item and share that information with the next set of test-takers (eg, the patient with the moustache has aortic stenosis). One option for avoiding this is to write multiple items for each piece of media.
- ▶ The richness of the patient description required: A long clinical vignette combined with media such as a video clip can provide a rich description of the patient that is more authentic to clinical practice, as it requires the students to interpret findings. However, this type of item would require more time for the examinee to explore the media before reviewing the options. Item writers should be aware of the trade-offs between the desired level of richness and the additional time or effort required on the part of the examinee.

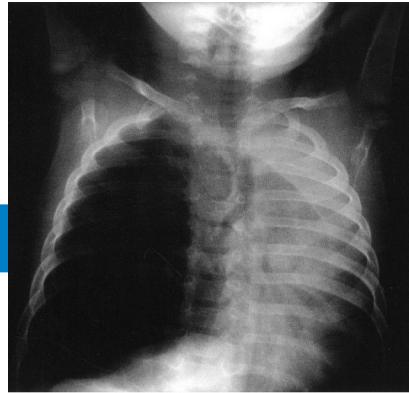
TYPES OF MEDIA

There are several types of media commonly used in health sciences examinations:

- ▶ Static images (eg, imaging results)
- ▶ Patient photographs
- ▶ Videos
- ▶ Interactive media (eg, avatars)
- ▶ Sound files (eg, heart or breath sounds)

Static Images

Radiographic studies, such as x-rays, CT scans, and MRIs, are common image types included in multiple-choice questions. Students may be asked to interpret the studies and decide on a diagnosis or management plan. Two examples are shown here:



Patient Photographs

Patient photographs add authenticity and provide a means of conveying information accurately and succinctly to the student. Two example photographs are shown here:



Videos

In general, videos can be useful to show physical examination findings as well as patient-provider interactions. For example, findings from a neurologic examination are much better shown than described. A screenshot of a video is shown below with an accompanying item. The video shows the resting tremor of Parkinson disease.



A 70-year-old man comes to the clinic because of difficulty writing during the past 3 months. He says that he fell once without injury this past week. He works as an accountant. He has a history of hypertension and hyperlipidemia. His medications include hydrochlorothiazide and atorvastatin. His vital signs are within normal limits. The remainder of the examination is remarkable only for the findings shown. Play the video to view the exam. Which of the following is the most likely diagnosis?



- A. Benign essential tremor
- B. Dementia, Alzheimer type
- C. Parkinson disease*
- D. Subclavian steal syndrome

Interactive Media and Sound Files

Media can also be interactive, requiring students to select areas to see and/or hear different examination findings that are similar to actual examinations. An example screenshot of an avatar simulating placement of a stethoscope with corresponding heart sound is shown here:



SELECTING MEDIA

In multiple-choice examinations, media should be purposefully selected to help the student answer the question; otherwise, it is simply extraneous information. Do not describe with text that which can be easily demonstrated in the media itself. In the example below, three similar items are shown with differing lead-ins and media; no media (Example A), a static image depicting heart rhythms or sounds (Example B), and an avatar simulating placement of the stethoscope (Example C). Other possibilities include showing both the ECG and the avatar or presenting the audio file of the corresponding heart sounds with or without a live patient video.

Consider the following stem for a cardiology multiple-choice question:

A 27-year-old man who is a US Army veteran comes to the office because of periodic dizziness, palpitations, and chest tightness during the past 3 weeks. The episodes occur when he remembers “the roadside bomb that took my friend.” He has had difficulty sleeping and drinks 1 pint of vodka daily to help with “nerves.” He has no documented medical history and takes no medications. Temperature is 36.7°C (98.1°F), pulse is 90/min, respirations are 20/min, and blood pressure is 128/80 mm Hg.

Below are three possible lead-ins and media selections for the above stem.

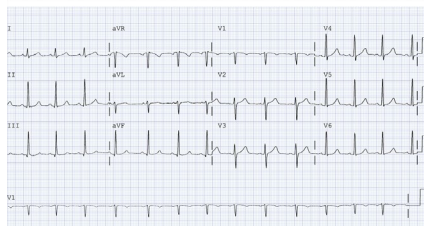
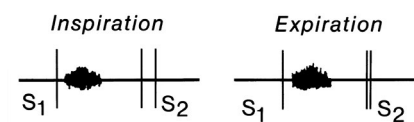
EXAMPLE A (no image)



Which of the following is the most likely finding on cardiac auscultation of this patient?

- A. Normal examination*
- B. Opening systolic snap
- C. S4 gallop
- D. S3 gallop
- E. Systolic flow murmur

EXAMPLE B (with ECG image)



An ECG is shown. Which of the following is the most likely finding on cardiac auscultation?

(same options as above)

EXAMPLE C (with avatar that allows auscultation of actual heart sounds through headphones)



An avatar is shown. Click the yellow circles to hear the cardiac examination. Which of the following is the most likely finding on cardiac auscultation?

(Same options as the preceding example)

CONTENT AREAS CONDUCTIVE TO THE USE OF MEDIA

Certain content areas lend themselves well to the use of media, such as:

- ▶ Dermatologic and musculoskeletal examination findings
- ▶ Cardiology (such as heart sounds)
- ▶ Neurologic examination findings
- ▶ Ethical and communication scenarios

Examples of two of these areas follow.

Dermatologic and Musculoskeletal Examination Findings

Dermatologic and musculoskeletal examination findings in particular benefit from the use of media. Showing findings, rather than describing the findings with text, simulates real clinical practice. Further, research has shown that response time is faster with the use of media compared with text for dermatologic findings. Consider the two following examples.

Example Item Using Text



A 79-year-old woman comes to the office 8 weeks after noticing a nontender nodule on the back of her left hand. She initially thought it was an insect bite, but it has grown in size over the past week. It bleeds when she picks at it. She has no history of serious illness. She lives in a retirement community in Texas and is an avid gardener. Examination of the dorsum of the left hand shows a 2-cm lesion that is well-demarcated, raised, and flesh-colored at the margins, with a necrotic center. Which of the following is the most appropriate next step in management?

- A. Cryotherapy
- B. Electrocautery ablation
- C. Excision of the lesion*
- D. Topical ketoconazole
- E. Observation

Example Item Using Media



A 79-year-old woman comes to the office 8 weeks after noticing a nontender nodule on the back of her left hand. She initially thought it was an insect bite, but it has grown in size over the past week. It bleeds when she picks at it. She has no history of serious illness. She lives in a retirement community in Texas and is an avid gardener. Examination of the dorsum of the left hand shows a 2-cm lesion. A photograph of the lesion is shown. Which of the following is the most appropriate next step in management?

(Same options as the preceding example)

Ethical and Communication Scenarios

Text-based multiple-choice questions related to ethical and communication scenarios tend to be easy for the student to answer because intonation and body language cannot be demonstrated well through text. Consider the next two examples.

Example Item Using Text



An 83-year-old woman is hospitalized for pneumonia and renal failure. She has a history of dementia, Alzheimer type, and resides in a nursing care facility. She has been offered but has refused dialysis. The patient has not designated a durable power of attorney, but she does have an advance directive that states, "No CPR, no intubation, no dialysis, and no surgery." The patient's niece, who is her closest relative, has a discussion with the provider about her aunt's refusal of treatment. She states, "Dialysis isn't aggressive, right? I have a neighbor who has been getting dialysis for 5 to 6 years! I want my aunt to get dialysis." Which of the following is the most appropriate provider response to the niece?

- A. "I can see you are upset. Would you like to speak with one of our chaplains or social workers?"
- B. "I know it's difficult, but you're going to have to accept that your aunt will not be getting dialysis or any other aggressive treatment. We would be breaking the law to treat her against her wishes."
- C. "I understand how you feel. Last year, I had a dear friend who died of cancer. It can be so hard to watch and not do anything."
- D. "I'm so sorry we can't do what you want. Let's discuss what we can do to make sure your aunt is comfortable."*
- E. "You do not need to worry; even though we can't do dialysis, we will try to make sure your aunt doesn't suffer. Do you have anyone to turn to for support right now?"

Example Item Using Media (screenshot from an approximately 30-second interaction)



An 83-year-old woman is hospitalized for pneumonia and renal failure. She has a history of dementia, Alzheimer type, and resides in a nursing care facility. She has been offered but has refused dialysis. The patient has not designated a durable power of attorney, but she does have an advance directive that states, “No CPR, no intubation, no dialysis, and no surgery.” The patient’s niece, who is her closest relative, has a discussion with the provider about her aunt’s refusal of treatment. Play the video to view the conversation. Which of the following is the most appropriate response to the niece?

(Same options as the preceding example)

ACQUIRING AND CREATING MEDIA

When determining new media needs, a subject matter expert group can be helpful as part of the process to oversee and monitor the acquisition process. This group can develop a list of diseases, conditions, and/or provider tasks and skills that are best illustrated with media. Once media are acquired, this group can develop exemplars to distribute for item-writing assignments. A good media image is one to which multiple test items can be written; this allows for the highest chance of the image being suitable for the exam and helps address the issue of memorability. It can also offset the cost of acquiring media. We also encourage diversity in portrayals of common conditions (eg, different skin tones and types). See “patient characteristics” section on page 42.

When acquiring media, two important considerations are patient confidentiality and metadata (ie, the information that accompanies and identifies each media image). If actual patient images or videos are used, it is important to maintain patient confidentiality. Ensure that neither the patient nor the institution can be identified from any clues in the media. For guidance, refer to your institution’s patient confidentiality policy and HIPAA guidelines (<http://www.hhs.gov/hipaa/for-professionals/index.html>).

Metadata is the identifying information that accompanies each piece of media. It is important to obtain as much metadata as possible about the media to help with indexing/searching and reuse in the future. Think about search terms and metadata that will help with identification of images that will be used more than once. It is advisable to set standards for media and copyright of media and to use a form to record as much metadata as possible during the acquisition stage. The following is a sample list of information to collect and record for media material:

- ▶ Administrative details
- ▶ Age of the patient
- ▶ Diagnosis
- ▶ Keywords
- ▶ Description of the test being performed
- ▶ Normal or abnormal results
- ▶ Descriptive file name
- ▶ Patient ID/name
- ▶ Indication patient signed a consent form
- ▶ In/out cut points for individual clips
- ▶ Whether the clip contains important audio

REMEMBER: YOUR MEDIA ARE ONLY AS GOOD AS THEIR METADATA!

Media have little value in item writing if they cannot be retrieved easily in search results.

Sources of Media

There are several sources from which to acquire media. The subject matter or content experts' personal libraries and/or patients are often a ready option, but the issues of patient confidentiality and memorability should be addressed. Furthermore, this is limited by the available patient population (eg, item writers may not have access to a patient with the symptoms that are best to show). A second option is to purchase existing media from vendors. This is the more expensive option, but it often allows the test developer to request specific content areas and types of media, along with specific instructions that will help decrease memorability. A third option is to create new media, either in-house or with a vendor. One example would be to record a specific set of videos for use in test items related to communication skills, using actors to portray both providers and patients.

When acquiring media, there are some guidelines to follow to help avoid technical issues. Select a specific format that will work with your exam software and verify that the media are in that format. Converting or editing files can be problematic, so it is better to have media created in the format you need rather than to have to convert a format. For static images, do not use media that is already embedded in other software (eg, Microsoft PowerPoint, Microsoft Word) or screenshots of previously published or widely available images. The more an image or video is manipulated, the more chance it has of losing its original quality; ideally media in an exam should be as high quality as possible.

With any media, the creation process should focus on protecting patient confidentiality (if real patients are used) and minimizing distractions for test-takers. A distraction is a specific type of cue that may remove the test-taker's focus from the important aspect of the video (eg, busy background, clothing with logos) and increase the time it takes for them to respond to the items, resulting in a more difficult item than intended. Distractions can also add to the memorability of the item. Even professional media content vendors are unlikely to be savvy about test development or to understand the impact that novelties and distractions can have on test-taker performance. The item writer and test developer should provide guidelines to help media content providers develop and provide videos that maintain patients' anonymity and minimize distractions.

Review the screenshot below taken from a video. List the things in the image that provide visual cues (distractions) for test-takers [hint: there are 12 visual cues in the image].



The following are the visual cues (ie, distractions) to the test-taker:

1. Setting (auditorium; could aid with memorability)
2. Electrical socket is specifically identifiable within the auditorium (could aid with memorability)
3. Clothes in the aisle
4. Patient's watch
5. Green shirt
6. Provider's ring
7. Blue and white striped shorts
8. Provider's watch
9. Provider's face/expression
10. Provider's shirt (white coat is preferable)
11. Provider wearing shorts
12. Bearded man in the background (mysterious; could aid with memorability)

TIPS FOR CREATING VIDEOS

DO:

- ▶ Use a plain background
- ▶ Avoid visual cues (eg, office equipment, paintings)
- ▶ Record in a well-lit room
- ▶ Have the patient wear plain clothes or a hospital gown with no logos
- ▶ Keep background and clothing consistent if there are multiple patients or exams
- ▶ Have the provider talk to the patient as they would in a normal exam
- ▶ Have the provider avoid using names when addressing the patient
- ▶ Leave the videos as you record them and provide instructions for editing
- ▶ Limit the video length to 30 seconds
- ▶ Provide a signed patient consent form

DON'T:

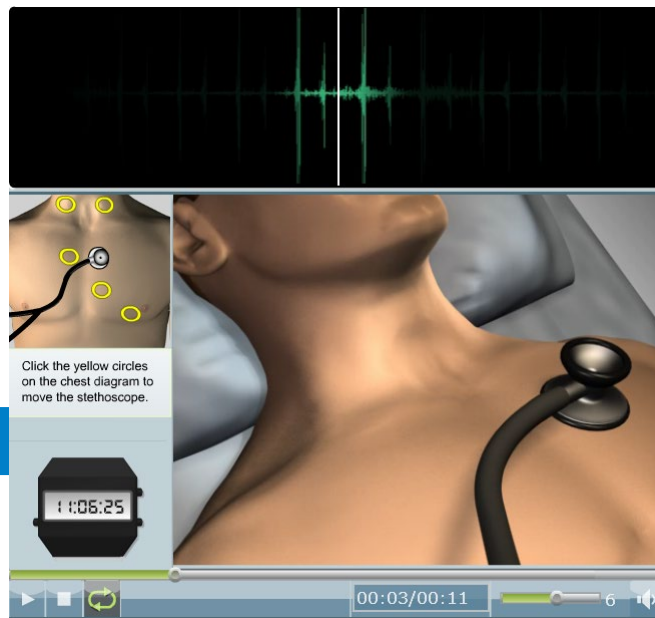
- ▶ Pick a brightly colored room, a room with identifiable pictures on the wall, or a room with unique (nonclinical) furniture
- ▶ Allow the patient or provider to wear brightly colored clothes, clothes with logos, or jewelry
- ▶ Show the faces of the provider(s) or patient(s) if it is not essential
- ▶ Explain everything in detail or add narration
- ▶ Add transitions (eg, fade-in, fade-out) to video files
- ▶ Resize or change the dimensions of the video

MEDIA ACCESSIBILITY

When using media in your tests, it is critical to keep accessibility in mind for examinees with disabilities. It is recommended to address issues of accessibility at the beginning of the media and item development process since addressing them later could result in additional complexity.

Media Accommodations for Hearing Impairment

Test takers who have hearing impairments may need captions or subtitles added to videos that include an audio component. Clinical-related audio such as heart sounds should be provided with a visual so that the candidate is able to interpret the sounds. In the heart sound example below, a visual waveform animation displays in sync with the heart sounds for test-takers with hearing impairment.



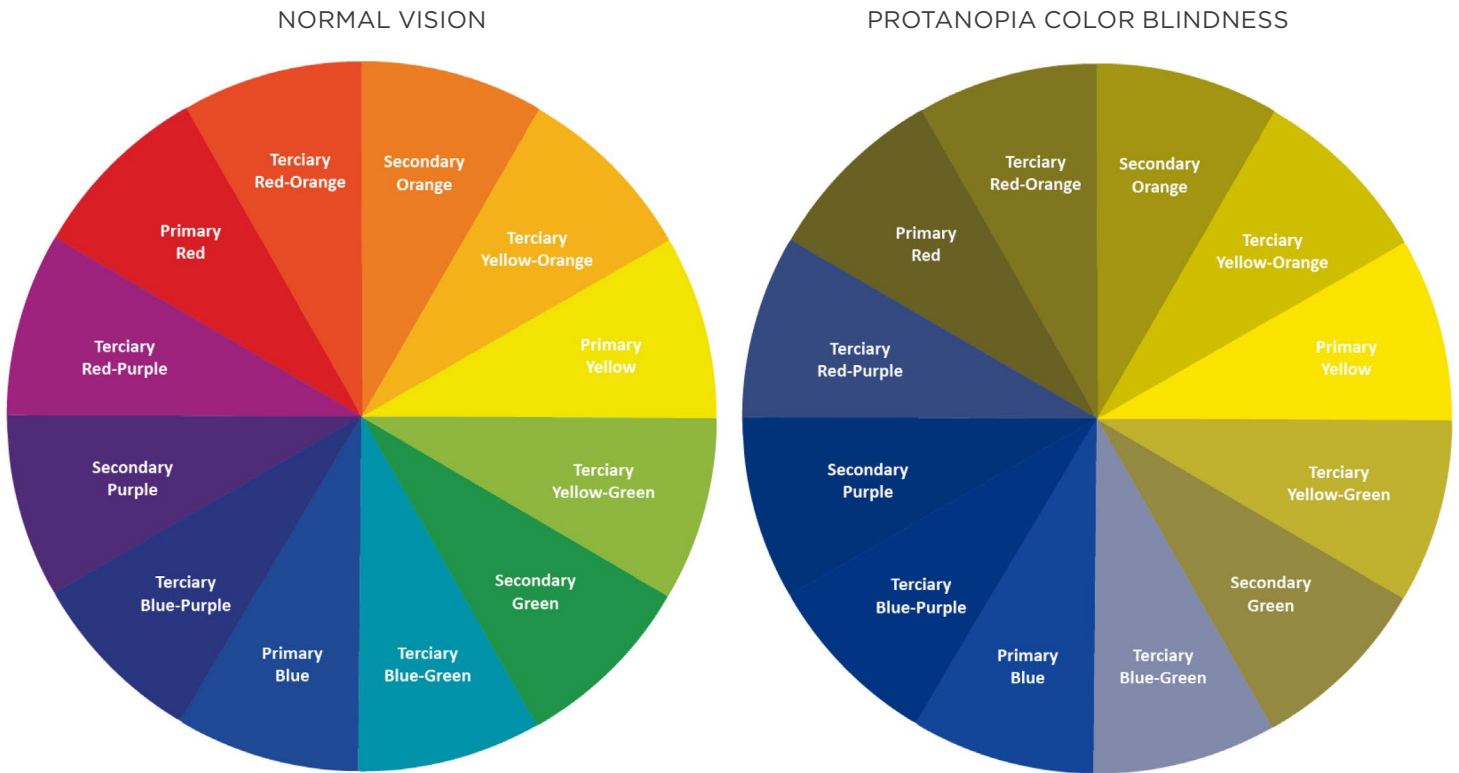
Media Accommodations for Visual Impairment

For test-takers with visual impairment, images should be provided with text descriptions where possible. The descriptions should be written in a way that do not provide an unfair cue or advantage as to the correct answer. Too much information can result in an advantage to candidates taking the accommodated version of the exam.

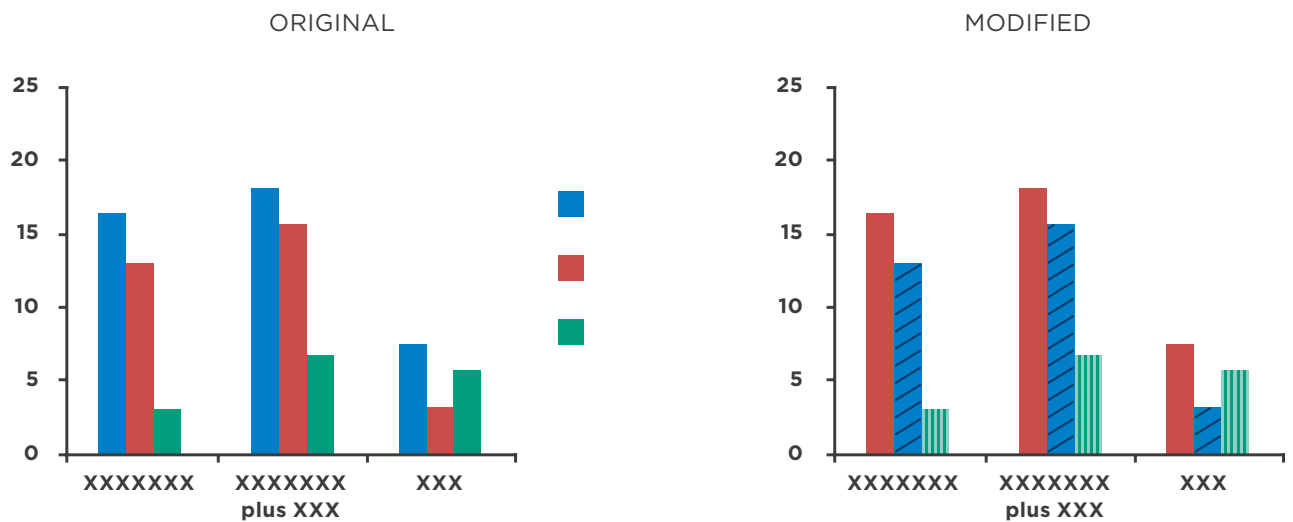
Color blindness is fairly common and should be a serious consideration when creating media for an examination. Charts, graphs, and highlighting can be impossible to interpret for test-takers who have color blindness. It is recommended that only black and shades of gray be used whenever possible. If using or designing graphics in which color distinctions are critical, it is helpful to keep the following in mind:

The most common type of color blindness is red-green color blindness. It is recommended that those two colors not be used together.

The following chart can be used to guide which colors will work together when designing graphics in which color distinctions are critical.



Patterns should be used wherever possible (eg, in bar graphs).



Sans serif fonts, such as Arial, are easier for candidates with visual impairments to read. Make sure that text size is no smaller than 8 pts or that the media has zoom-in functionality so that examinees with visual impairments can read the text.

SECTION 3:

APPENDICES



SM

APPENDIX A: A QUICK REFERENCE GUIDE TO APPROACHING ITEM WRITING

Getting Started

- ▶ Consider the curriculum: What needs to be covered? At what level of learner?
- ▶ Consider classic and more common presentations you've encountered in your own practice/setting as a starting point
- ▶ Think of what you would want your test population to recognize (or not miss)

"...just saw a classic case of renal artery stenosis last week..."

The Testing Point

- ▶ Think of what you want to test (the "testing point") as you start writing

Principles of therapy in renal artery stenosis...

The Site of Care

- ▶ Where do you envision the site of care for the case?

Would I treat this in the clinic? Hospital?

The Case

- ▶ Start outlining the case that frames your testing point (see Chapter 6: Testing Application of Foundational [Basic] and Clinical Knowledge)
- ▶ View sample items as a guide to the style (see Chapter 6 under "Structuring Items to Fit Task Competencies")
- ▶ Consider if an image or other media could work as well as or better than a text description (keeping in mind that the question should not be answerable based on the image alone) (see Chapter 7: Using Media as Part of Clinical Vignettes)

"A 65-year-old man comes to the clinic because of a 2-week history of swelling of his ankles and feet....He has an 8-year history of type 2 diabetes mellitus and hyperlipidemia. Medications are...."

The Lead-in

- ▶ Ask: What will you have the provider do?
- ▶ See Appendix B: Sample lead-ins; focus on the section that matches your testing point objective (management, diagnosis, etc.)

Management lead in: "Which of the following is the most appropriate pharmacotherapy at this time?"

The Option Set

- ▶ Aim to have enough information in your vignette to support the key (correct answer) and reasonably link to your distractors (wrong answers)
- ▶ Ensure your option set is free of flaws (see Chapter 3: Technical Item Flaws)

A. Drug
B. Drug
C. Drug
D. Drug
E. Drug

APPENDIX B: SAMPLE LEAD-INS BASED ON TASK COMPETENCIES

Medical Knowledge: Applying Foundational Science Concepts

Foundational (basic) science comprises items that require understanding and application of basic science principles to answer the question. Foundational science items should not be answerable based simply on clinical knowledge alone or on pattern recognition (eg, providing a list of symptoms and asking what drug to prescribe). These items should require clinical knowledge and also knowledge of one or more foundational science principles that would have likely been learned in preclinical education and hopefully reinforced during clinical rotations.

Patient Care: Diagnosis—Causes and Mechanisms

Identifies the cause/infectious agent or predisposing factor(s) or, given an effect, determines the cause.

- ▶ Which of the following pathogens is the most likely cause of this patient's condition?
- ▶ Which of the following is the most likely infectious agent?
- ▶ This patient most likely acquired the infectious agent via which of the following modes of transmission?
- ▶ This patient most likely has a defect in which of the following?
- ▶ Which of the following is the most likely cause/mechanism of this effect?

Identifies the underlying processes/pathways that account for, or contribute to, the expression or resolution of a given condition.

- ▶ Which of the following is the most likely underlying cause of this patient's condition?
- ▶ Which of the following is the most likely explanation for this patient's condition?
- ▶ Which of the following cell types most likely played a primary role in the development of this lesion?
- ▶ Which of the following immune system mediators plays a critical role in the pathogenesis of this patient's current condition?
- ▶ This patient most likely has a deficiency in which of the following enzymes?
- ▶ Which of the following cytokines is the most likely cause of this condition?
- ▶ Which of the following processes is most likely impaired in this patient?

Recognizes or evaluates given clinical or physical findings to identify the underlying anatomic structure or physical location.

- ▶ The most likely cause of the findings in this patient is damage to which of the following structures?
- ▶ Which of the following structures is at greatest risk for damage during this procedure?
- ▶ Which of the following nerves is most likely carrying the sensation for this patient's pain?
- ▶ The most likely cause of these findings is dysfunction of which of the following structures?
- ▶ Which of the following developmental abnormalities is the most likely cause of the findings in this patient?

Recognizes the mechanisms of action of various drugs; selects from an option set list of drugs based on mechanism of action.

- ▶ Which of the following is the most likely mechanism of the beneficial effect of this drug?
- ▶ Which of the following is the most appropriate management? (response options would be drug classes or mechanisms of action)
- ▶ Which of the following is the most likely mechanism of action of the therapeutic effect of this drug?
- ▶ The most appropriate medication for this patient will have which of the following mechanisms of action?

Patient Care: Diagnosis—Obtaining and Predicting History and Physical Examination

Knows signs/symptoms of selected disorders. Response options are signs and symptoms. The item asks which signs and symptoms are characteristic of the patient's condition. Typically used when patient presents with the condition.

- ▶ Which of the following signs/symptoms is most consistent with the underlying diagnosis in this patient?

Knows individual's risk factors for development of condition. Given current symptoms in presented history, identifies pertinent factor in the history. Typically used when patient presents with the condition.

- ▶ Which of the following factors in this patient's history most increased the risk for developing this condition?

Given a specific problem, knows what to ask to obtain pertinent additional history. The response options should not be referenced in the vignette and should not include details that would be obtained during initial history-taking. If asking about information that was already obtained and is mentioned in the vignette, use the following lead-in.

- ▶ It is most appropriate to obtain specific additional history regarding which of the following?

Predicts the most likely additional physical finding; selects either the finding itself or the appropriate examination technique that would result in the finding. The options are findings or directed physical examination techniques.

- ▶ The remainder of the physical examination is most likely to show which of the following? (ensure all options are portions of the physical examination that would not yet have occurred in the patient scenario)
- ▶ It is most appropriate to direct the physical examination toward which of the following? (sample options: “Auscultation of the lungs,” “Palpation of the abdomen,” “Rotation of hip joints”)
- ▶ Which of the following signs or symptoms is most indicative of the need for further studies?

Patient Care: Diagnosis—Selecting and Interpreting Laboratory and Diagnostic Studies

Selects most appropriate laboratory or diagnostic study, including neuropsychiatric testing, or study most likely to establish/confirm the diagnosis. Options can include “No further testing is indicated.”

- ▶ Which of the following is the most appropriate diagnostic study at this time?
- ▶ Which of the following is the most appropriate initial diagnostic study?
- ▶ Which of the following is the most appropriate next step in evaluation?
- ▶ Which of the following studies is most likely to establish a diagnosis?
- ▶ Which of the following laboratory studies is most likely to confirm the diagnosis?
- ▶ Which of the following studies is most appropriate to order for follow-up?
- ▶ Which of the following studies is most appropriate to obtain periodically to monitor the patient’s long-term care?

Interprets laboratory or other study findings. Response options are interpretations of the laboratory/diagnostic data.

- ▶ Which of the following is the most likely explanation for these laboratory findings?
- ▶ Based on these findings, this patient’s condition is most likely attributable to which of the following?

Predicts the most likely laboratory or diagnostic study result. Response options are clinical studies or predicted study results.

- ▶ Results of which of the following laboratory studies are most likely to be abnormal in this patient?
- ▶ Measurement of serum electrolyte concentrations is most likely to show which of the following?
- ▶ An x-ray of the _____ is most likely to show which of the following?
- ▶ Arterial blood gas analysis is most likely to show which of the following sets of findings?

Selects most appropriate laboratory or diagnostic study after change in patient status.

- ▶ Prior to changing this patient's therapy, which of the following is the most appropriate diagnostic study?

Patient Care: Diagnosis—Formulating the Diagnosis

Selects the most likely diagnosis.

- ▶ Which of the following is the most likely diagnosis?
- ▶ Which of the following is the most likely working diagnosis?
- ▶ Which of the following best explains these findings? (options would be diagnoses; correct answer could be "Normal finding(s)")
- ▶ Which of the following infectious agents is the most likely cause of this patient's pneumonia?

Patient Care: Diagnosis—Determining Prognosis/Outcome

Recognizes factors in the history, or physical or laboratory study findings that affect patient prognosis or outcome, or that determine therapy.

- ▶ Which of the following factors in this patient's history most strongly indicates a poor/good prognosis?
- ▶ Which of the following factors is most critical in determining this patient's ability to remain at home?

Interprets laboratory or other diagnostic study results and identifies current/future status of patient.

- ▶ Based on these findings, this patient is most likely to develop which of the following?
- ▶ Based on these findings, this patient is most likely to develop which of the following during _____ (period of time)?

Recognizes associated conditions of a disease, including complications, or indicators of potential complications.

- ▶ Which of the following is the most likely complication of this patient's current condition?
- ▶ Without treatment, which of the following is most likely to develop in this patient?

Recognizes characteristics of disease relating to natural history or course of disease, including progression, severity, duration, and transmission of disease for a specific patient.

- ▶ Which of the following is the most likely clinical course for this patient?
- ▶ It is most appropriate to inform this patient of which of the following risk factors?

Patient Care: Management—Health Maintenance and Disease Prevention

Knows risk factors for conditions amenable to prevention or detection in an asymptomatic patient or knows the potential condition itself.

- ▶ Which of the following is the strongest predisposing factor in this patient for developing a chronic condition?
- ▶ It is most appropriate to counsel this patient that he/she is at greatest risk for which of the following?
- ▶ If untreated, this patient is at greatest risk for which of the following disorders?

Knows pertinent incidence statistics and identifies patient groups at risk; knows incidence of symptomless/dangerous disorders among various groups. Response options compare patient's risk factors for disease with those of the general population.

- ▶ Which of the following factors is most appropriate to consider in assessing the need for additional screening in this patient population?

Knows common screening tests for conditions amenable to prevention or detection in an asymptomatic patient or population.

- ▶ In addition to a [screening test] annually, which of the following screening studies is most appropriate?
- ▶ Which of the following is the most appropriate screening test for this patient at this time?
- ▶ At this time, which of the following is the most appropriate next step in the evaluation of this patient? (response options would be screening tests)

Selects appropriate preventive agent or technique (eg, contraception, vaccines, vitamins). Knows timing of vaccinations.

- ▶ Which of the following is the most appropriate recommendation for vaccination?
- ▶ [To reduce the likelihood of recurrence,] it is most appropriate to (prescribe/administer) which of the following?
- ▶ Which of the following is the most appropriate vaccine to administer at this time?

Knows appropriate counseling (and reassuring, comforting) of patient or family regarding current and future problems, including risk factors related to present encounter. The response options focus on features and course of disease as they relate to a specific patient.

- ▶ It is most appropriate to counsel this patient regarding which of the following?
- ▶ Which of the following is the most appropriate management to prevent the spread of this patient's illness?

Educates patients on screening, health maintenance, and self-care options such as nutrition, weight loss, breast self-examinations, home blood pressure monitoring, or breast-feeding. The response options are a variety of patient actions.

- ▶ Which of the following is the most appropriate recommendation?
- ▶ It is most appropriate to advise this patient to do which of the following?

Patient Care: Management—Selecting and Monitoring Pharmacotherapy

Selects most appropriate pharmacotherapy. Response options are all drugs and can include “No pharmacotherapy at this time.” Drugs include substances such as vitamins and supplements.

- ▶ Which of the following is the most appropriate pharmacotherapy at this time?
- ▶ The most appropriate next step is to administer which of the following?

Assesses patient adherence to treatment regimen, recognizes techniques to increase adherence to or understanding of the disease state, and knows how adherence may be affected by providing instructions with therapy. Options can include “No further testing is indicated.”

- ▶ To assess adherence and therapeutic efficacy in this patient, which of the following studies is most appropriate to order after initiation of therapy?
- ▶ Which of the following methods is most appropriate to assess adherence to treatment in this patient?

Recognizes factors that alter drug requirements for a patient, such as disease, age, pregnancy, BMI, renal failure, liver disease, or gender. Response options are factors about the patient that affect the choice of a drug regimen.

- ▶ Which of the following variables is most appropriate to consider in determining the appropriate dose of medication for this patient?
- ▶ Which of the following factors is most likely to influence therapy for this patient?
- ▶ Which of the following is most appropriate to consider before selecting pharmacotherapy?

Knows adverse effects of various drugs or recognizes signs and symptoms of drug (and drug-drug) interactions resulting from polypharmacy in the therapeutic regimen and knows steps to prevent polypharmacy including laboratory studies to monitor pharmacotherapy. Vignette includes description of the simultaneous use of drugs prescribed by another provider, over-the-counter drugs, prescribed opioids and other Schedule IV medications taken illegally or in greater than prescribed doses, illegal opioids, alcohol, and certain foods.

- ▶ Which of the following is the most likely cause of this patient's symptoms?
- ▶ Which of the following is the most likely complication of the addition of this medication?
- ▶ Which of the following is the most likely explanation for this patient's current condition?
- ▶ The most likely cause of this patient's condition is interaction between which of the following drugs?

Knows contraindications of various medications.

- ▶ Which of the following medications is contraindicated in this patient?
- ▶ Which of the following medications is most likely to increase the risk for development/progression of _____ (diagnosis) in this patient?

Knows modifications of a therapeutic regimen within the context of continuing care.

- ▶ Which of the following is the most appropriate next step in pharmacotherapy?
- ▶ Which of the following is the most appropriate change/modification in this patient's pharmacotherapy?

Knows appropriate monitoring to evaluate effectiveness of pharmacotherapy or to monitor for the adverse effects of pharmacotherapy in a patient who has not had a recurrence or progression of disease.

- ▶ Which of the following studies is most appropriate to monitor the effectiveness of therapy in this patient?

Patient Care: Management—Clinical Interventions/Treatments

Knows most appropriate management of selected conditions, including recognizing use/misuse of medications, alcohol, or other substances. Response options would be a list of management steps.

- ▶ Which of the following is the most appropriate next step in management?
- ▶ Which of the following is the most appropriate initial management/recommendation?

Knows immediate management or priority in management, specifically in emergency or acute cases. This objective is most appropriate in life-threatening emergencies or cases of potential organ failure.

- ▶ Which of the following is the most appropriate immediate/initial/next step in management?
- ▶ Which of the following is the priority in management?
- ▶ Which of the following is the most critical factor in formulating a management plan for this patient?

Knows most appropriate follow-up or monitoring approach regarding the management plan.

- ▶ Which of the following is the most appropriate monitoring/follow-up plan?

Knows current/short-term management of patients.

- ▶ Which of the following is the most appropriate next step following treatment?
- ▶ Which of the following is the most appropriate next step in monitoring this patient?

Evaluates severity of patient condition in terms of need for referral for surgical treatments/procedures versus other nonsurgical options.

- ▶ Which of the following findings in this patient indicates the need for surgical intervention/intubation/transplantation/admission to another department?

Knows appropriate surgical management. The response options are all surgical procedures.

- ▶ Which of the following is the most appropriate surgical management?

Knows preoperative/postoperative or procedural management.

- ▶ Which of the following is the most appropriate postoperative management?
- ▶ Which of the following is the most appropriate preoperative preparation?
- ▶ Prior to the procedure (or specify the procedure), it is appropriate to first obtain/do which of the following?

Knows indications for admission to the hospital or to other appropriate setting. Knows appropriate nonhospital health care settings, such as a nursing care facility, hospice care, or at-home care with assistance of health aide.

- ▶ Placement in a/an _____ is recommended based on which of the following factors?
- ▶ To which of the following inpatient facilities is it most appropriate to transfer this patient?

Knows most appropriate discharge planning.

- ▶ In discussing discharge plans with this patient, it is most appropriate to advise him/her of which of the following?
- ▶ Before this patient is discharged, it is most appropriate to counsel him/her regarding which of the following?
- ▶ Which of the following is the most appropriate goal for follow-up?

Knows components of rehabilitation program, such as prostheses, psychosocial factors, or motor dysfunction. The response options are rehabilitation management steps.

- ▶ Which of the following components of his/her overall care is most appropriate to consider?

Knows appropriate use and procedures regarding hospice care.

- ▶ Which of the following is the most appropriate step regarding hospice care for this patient?
- ▶ Which of the following is the most appropriate next step? (hospice referral is correct answer)

Educates patient or family regarding self-care, such as breast-feeding, or at-home blood pressure measurement and glucose monitoring. The response options can be a variety of patient actions.

- ▶ Which of the following is the most appropriate recommendation?
- ▶ It is most appropriate to advise this patient to do which of the following?

Knows relevant roles of allied health personnel.

- ▶ The most appropriate next step is to arrange consultation with which of the following?
- ▶ It is most appropriate to refer this patient to which of the following?

Patient Care: Management—Selecting Clinical Interventions (Mixed Management)

Selects most appropriate option from set of mixed management options (eg, mix of diagnostic studies, pharmacotherapy, procedures, or no intervention at this time, observation, referral).

- ▶ Which of the following is the most appropriate next step?
- ▶ Which of the following is the most appropriate initial step in management?

Patient Care: Management—Monitoring/Surveillance for Disease Recurrence or Progression

Knows the indications for surveillance for recurrence or progression of disease following treatment.

- ▶ Which of the following is the most appropriate annual monitoring study?

Knows how to monitor a chronic disease in a stable patient where a change in patient status might indicate a need to change therapy.

- ▶ Which of the following is the most appropriate diagnostic study at this time?

Knows most appropriate long-term treatment or management goals, including continued treatment of a known patient. In a patient with a chronic condition, knows preventive medicine.

- ▶ Which of the following is the most appropriate long-term management?
- ▶ It is most appropriate to advise this patient of which of the following long-term management goals?

Communication and Interpersonal Skills

- ▶ Which of the following is the most appropriate opening remark to this patient?
- ▶ Which of the following is the most appropriate response by the provider?
- ▶ Which of the following statements by the provider is most appropriate to...?

Professionalism and Legal/Ethical Issues

Knows the guidelines for obtaining informed consent for treatment including those for children and adolescents, third-party permission, and emergent situations.

- ▶ Which of the following is the most appropriate method to facilitate patient informed consent?

Recognizes need for third-party permission for treatment in medical emergencies.

- ▶ In requesting an autopsy, consent must be given by which of the following individuals?

Knows guidelines for treatment of minors with/without notification of parents.

- ▶ Regarding obtaining consent for treatment today, which of the following is the most accurate conclusion for providing services to this patient?
- ▶ Until the parents can be reached, which of the following is the most appropriate management?
- ▶ Before examining this patient, informed consent needs to be obtained from which of the following individuals?
- ▶ Which of the following is the most appropriate response to the request for services for this child/adolescent?

Knows definitions of competence and sanity.

- ▶ Determination of this patient's competency to make decisions should be based on which of the following factors?
- ▶ Which of the following is the most significant indication that this patient may not have sufficient capacity for informed consent or refusal?
- ▶ Which of the following factors or findings renders this patient incompetent to make health care decisions on her/his own behalf?
- ▶ The information that is most decisive in determining this patient's mental capacity to refuse treatment is which of the following?

Knows the guidelines for involuntary admission (eg, third-party permission, court order).

- ▶ Which of the following is the most appropriate next step? (answer is involuntary admission)
- ▶ In considering commitment of this patient to _____, which of the following is the most appropriate next step?

Knows guidelines for such things as confidentiality of medical records (eg, regarding patient's relatives, employer, insurance/legal agents), boundaries, privacy, and truth-telling.

- ▶ Which of the following is the most accurate statement concerning the confidentiality of this patient's medical record?

Knows guidelines for provider/patient relationship.

- ▶ Regarding this colleague's behavior, which of the following is the most appropriate advice to this patient?

Assesses degree of disclosure to terminally ill patients.

- ▶ Which of the following is the primary consideration regarding informing this patient of his/her condition/prognosis?

Recognizes patient's right to refuse treatment or testing (patient autonomy); knows issues of advance directives and living wills.

- ▶ Which of the following is the most appropriate advice to the family regarding their wishes for this patient?
- ▶ Given your knowledge of this patient and her past wishes, which of the following is the most appropriate recommendation?
- ▶ In reaching an opinion about whether do-not-attempt-resuscitation status should be ordered, which of the following information has the highest priority?
- ▶ The best method of ensuring that this patient's wishes will be honored is to do which of the following?

Assesses quality-of-life decisions (especially in elder patients).

- ▶ In considering treatment options for this patient, which of the following is the most compelling consideration?
- ▶ Which of the following is the most appropriate next step regarding this patient's end-of-life needs?

Knows appropriate prescriptive practices; knows appropriate use of opioids in terminally ill patients.

- ▶ Which of the following is the most effective intervention to minimize this patient's pain?

Knows definition of and legal issues regarding brain death.

- ▶ Which of the following is the most critical ethical consideration in deciding whether to withdraw life support in this patient?
- ▶ Which of the following is the most accurate statement regarding the physician's decision to discontinue life-sustaining interventions in this patient?
- ▶ In order for a clinical diagnosis of brain death to be made in this child, documentation of which of the following is needed?

Knows management of terminally ill patients related to treating chronic pain, and recognizes patient's expression of fear of pain, injury, or death; knows how to comfort patient or family during crisis such as trauma or death.

- ▶ Which of the following is most appropriate to manage this patient's pain?
- ▶ Which of the following is the most appropriate advice for this patient regarding his/her pain?
- ▶ Which of the following is the most likely underlying cause of this change in behavior?
- ▶ In addressing this patient's (fear), which of the following is the most appropriate counseling?

Knows guidelines for reporting findings to proper authorities, such as social services, police, medical society, or coroner.

- ▶ Which of the following is the most appropriate action in patient care?

Knows Good Samaritan laws.

- ▶ Regarding your colleague's liability associated with this patient's actions at the scene of the collision, which of the following is the most accurate conclusion?

Recognizes provider error and negligence.

- ▶ With regard to the possibility that this situation represents medical malpractice, it is most critical to do which of the following?
- ▶ The most appropriate response to this situation is to do which of the following?
- ▶ After you document the error, the most appropriate management is to do which of the following?
- ▶ Which of the following is the most appropriate response to this allegation?

Recognizes and deals appropriately with impaired providers.

- ▶ The most appropriate response to the licensing board is to recommend which of the following?
- ▶ Which of the following is the most appropriate action? (where options describe dealing with impaired colleague)

Systems-based Practice and Patient Safety

Understands basic concepts and terminology, principles, and application of quality improvement science and outcome analysis.

- ▶ Which of the following is the most appropriate description of the deviation from official procedure?

Recognizes and optimizes human and environmental factors such as workplace design, standardization, and processes.

- ▶ Which of the following is the most appropriate next step in designing a standard process for...?
- ▶ Which of the following is most likely to improve patient satisfaction?
- ▶ Which of the following strategies is most likely to achieve this goal?
- ▶ Which of the following is the most appropriate next step by this hospital to improve its system of care?
- ▶ Which of the following is the most appropriate initial recommendation by the task force?
- ▶ Which of the following is most likely to improve outcomes in this situation?
- ▶ Which of the following is most likely to decrease morbidity/mortality in this situation?

Understands the role and characteristics of teams and communication strategies.

- ▶ Which of the following is most appropriate to ensure the success of this project?
- ▶ Which of the following actions is most likely to improve communication within this health care team?

Anticipates, recognizes, analyzes, and mitigates risk (sources of error).

- ▶ Which of the following is the most appropriate method to prevent/reduce the risk of transmission of this infection?
- ▶ Which of the following is most likely to prevent recurrence of this type of error?
- ▶ Which of the following is the most likely cause of the error?
- ▶ Which of the following is most likely to decrease the likelihood of this error happening again?
- ▶ Which of the following is the most appropriate next step?

Evaluates, reports, and responds to near-misses and system errors.

- ▶ Which of the following is the most appropriate action by the hospital staff immediately following the incident?

Practice-based Learning—Applied Biostatistics and Clinical Epidemiology

Understands and can apply principles of epidemiology and population health, including health status indicators, outbreak investigation, points of intervention.

- ▶ Which of the following is the most likely effect on estimates of disease incidence and prevalence?
- ▶ Which of the following is the annual incidence of _____ in this study?

Understands and can apply principles of study design/flaws, such as bias and confounding, and methods to address these flaws; understands and can apply statistical principles.

- ▶ The most likely cause of the study results is an error related to which of the following?
- ▶ Which of the following features of this study is of greatest potential concern?
- ▶ Which of the following potential flaws is most likely to invalidate this study?
- ▶ Which of the following best describes this study design?
- ▶ Which of the following is the major advantage of this study design?

Understands and can apply principles of screening and other tests (eg, sensitivity, specificity, predictive value).

- ▶ Which of the following is the most likely effect on sensitivity and specificity?
- ▶ Which of the following is the most likely effect on the predictive value?

- ▶ If the prevalence of the disease is increased to ___%, which of the following would be the most likely outcome?
- ▶ According to these results, which of the following represents the sensitivity of _____ for detecting _____ within this population?
- ▶ According to these results, which of the following is closest to the predictive value of a positive test result?
- ▶ If _____ were decreased to _____, which of the following would be the most likely result?
- ▶ Changing the screening population is most likely to have which of the following effects on this test?
- ▶ In determining the appropriate diagnostic test for this patient, which of the following characteristics of the test is most appropriate to consider?
- ▶ Which of the following is the most appropriate conclusion regarding the test?
- ▶ Which of the following combinations of sensitivity and specificity would be characteristic of the most appropriate confirmatory/screening test?

Understands use and interpretation of statistical principles and measures of association.

- ▶ Which of the following conclusions can be drawn from these data?
- ▶ Which of the following is the most appropriate conclusion about these data?
- ▶ Which of the following is the most accurate interpretation of these data?
- ▶ Based on these additional data, it is most appropriate to conclude which of the following?
- ▶ Which of the following conclusions is most strongly justified based on this information/study?
- ▶ Which of the following is the most accurate interpretation of the author's conclusion regarding these study results?
- ▶ Which of the following is the most likely explanation for this finding?
- ▶ Which of the following represent the results of a meta-analysis?
- ▶ In determining the validity of the meta-analysis, which of the following is the most appropriate factor to consider?
- ▶ Which of the following is the best interpretation of this confidence interval?
- ▶ Compared with _____, which of the following is the relative risk for _____ in _____?
- ▶ Which of the following is the estimated odds ratio of _____ in _____ compared with _____?
- ▶ Which of the following is the best estimate of the relative risk of _____ for those with _____ compared with those with _____?
- ▶ Which of the following is the relative risk for _____ 5 years following _____?
- ▶ Which of the following is the attributable (excess) risk per _____ patients for development of _____ 5 years following _____?
- ▶ Which of the following variables is measured on a nominal scale?
- ▶ Which of the following is the mean (or mode or standard deviation) in the sample shown in the graph?
- ▶ Assuming a normal (Gaussian) distribution, which of the following best represents the median _____ for this group?

APPENDIX C: NBME RETIRED ITEM FORMATS

In order to move forward with innovations in assessment, it is necessary to look back on item types previously used on NBME exams and better understand item content or format flaws that can negatively impact the measurement of test-taker ability. Multiple item types were used on NBME exams beginning in the mid-1980s in order to provide sufficient variety for the relatively lengthy examinations, with predominant item types: A-type, B-type, C-type, and K-type items, along with the lesser used D-type, H-type, I-type, and R-type items. The current USMLE Step examinations now include only one of these item types, the A-type item, along with the more recently developed F-type and G-type items. The currently used NBME item types are summarized below:

- ▶ A-type: Standard one-best-answer item type
- ▶ F-type: Items grouped into sets around specific content, where test-takers cannot return to previously seen items in the set
- ▶ G-type: Items grouped into sets around specific content, where test-takers can return to previously seen items in the set

Limiting the number of acceptable item types that appear on an exam provides several advantages. First, it allows item authors to spend less time thinking about item types and more time concentrating on the most important aspects of item writing, such as clinical decision-making to test reasoning and problem-solving rather than recall. Second, it reinforces the standardized nature of the items within test forms, thus simplifying the process for test-takers and supporting the validity evidence of the exams. Third, it allows for more efficient production, editing, and approval of items.

Examples of retired item types and their flaws are described on the following pages, in alphabetic order by the letter used to designate item type. Please note that while these item types are no longer used on NBME exams, they may still be of potential use in the classroom or on course exams, but item writers should carefully review the potential flaws in order to properly evaluate the advantages and disadvantages of using each item type.

IN ORDER TO MOVE FORWARD WITH INNOVATIONS IN ASSESSMENT, IT IS NECESSARY TO LOOK BACK ON ITEM TYPES PREVIOUSLY USED AND BETTER UNDERSTAND ITEM CONTENT OR FORMAT FLAWS THAT MAY IMPACT THE MEASUREMENT OF TEST-TAKER ABILITY.

C-TYPE ITEMS

Noted Flaws

C-type items are similar to B-type items in appearance but are actually multiple true-false. The primary problem with C-types was in test-taker effort to conclude to what extent something had to be “true” to be selected. Using the example below, if one of the numbered options was associated with both A and B, but was more strongly associated with A, the test-taker would have to decide whether an appropriate response was A only or both A and B. These judgments were less about the medical knowledge and more about the test-taker having to guess the item writer’s intent.

Sample C-type item



DIRECTIONS: Each set of matching questions in this section consists of a list of four lettered options followed by several numbered items. For each numbered item, select the ONE lettered option that is most closely associated with it. Each lettered option may be selected once, more than once, or not at all.

- A. *Plasmodium vivax* malaria
 - B. *Plasmodium falciparum* malaria
 - C. Both
 - D. Neither
-
- 1. A combination of primaquine and chloroquine is treatment of choice for acute attack
 - 2. Clinical attacks suppressed by administration of chloroquine once a week while in treatment
 - 3. Infection prevented by administration of chloroquine once a week
 - 4. Permanently cured by administration of chloroquine

NOTES

D-TYPE ITEMS

Noted Flaws

D-type items are complex matching sets in which each item consists of three functional disturbances (designated by a letter) and five situations (in a numbered list). It was believed that these items required discriminatory understanding of a number of similar factors. However, D-type items were difficult to write, and the directions tended to be confusing to test-takers. In addition, these items did a poor job of discriminating among test-takers' level of knowledge.

Sample D-type item



DIRECTIONS: There are two responses to be made to each of the following questions. There are three lettered categories. Exactly four of the five numbered items are related in some way to ONE of these categories. (1) Choose the letter of the category in which these four items belong. (2) Then choose the number of the item that does NOT belong in the same category with the other four.

- A. Eosinophilia of diagnostic significance
 - B. Lymphocytosis of diagnostic significance
 - C. Plasmacytosis of diagnostic significance
-
- 1. Hodgkin disease
 - 2. Löffler syndrome
 - 3. Multiple myeloma
 - 4. Schistosomiasis
 - 5. Trichinosis

H-TYPE ITEMS

Noted Flaws

H-type items consist of paired statements describing two entities to compare in a quantitative sense. The test-taker was directed to select A if A was greater than B; B if B was greater than A; and C if the two were approximately equal. The H-type item was believed to be useful for those instances where recall of quantitative information is important, but it was challenging for test-takers to decide how great the difference needed to be in order to be relevant.

Sample H-type item



DIRECTIONS: The following paired statements describe two entities that are to be compared in a quantitative sense. For each numbered statement, choose

- A if (A) is greater than (B)
 - B if (B) is greater than (A)
 - C if the two are equal or very nearly equal
-
- 1. (A) The usual therapeutic dose of epinephrine
(B) The usual therapeutic dose of ephedrine
 - 2. (A) Life expectancy with glioblastoma of the occipital lobe
(B) Life expectancy with glioblastoma of the frontal lobe

I-TYPE ITEMS

Noted Flaws

The I-type item is similar to the H-type item. It consists of pairs of phrases that describe conditions or quantities that might vary in relation to each other. I-type items had two notable flaws. First, there were fewer options than in other item types, so there was an increased chance for the test-taker to guess the correct answer, which affected both item difficulty and discrimination. Second, the items tended to focus on minor details rather than on more relevant scientific concepts.

Sample I-type item



DIRECTIONS: Each of the following pairs of phrases describes conditions or quantities that may or may not be related. For each numbered statement, choose

- A if increase in the first is accompanied by increase in the second or if decrease in the first is accompanied by decrease in the second
 - B if increase in the first is accompanied by decrease in the second or if decrease in the first is accompanied by increase in the second
 - C if changes in the first are not necessarily accompanied by changes in the second
1. (A) Urine volume
(B) Urine specific gravity
 2. (A) Plasma protein concentration
(B) Colloid osmotic pressure of plasma

K-TYPE ITEMS

Noted Flaws

K-type items are multiple true-false sets that were once a common format at NBME. Because the items could include only absolutely true or false facts, K-type items could not be used to assess clinical judgment except in comparisons (eg, "Drug X is better than Drug Y in treating disease K"). Thus, they could appear too complicated and require the test-taker to constantly keep the answer code in mind. In addition, the possible response combinations introduced a cueing effect that decreased item discrimination.

Sample K-type item



DIRECTIONS SUMMARIZED

A	B	C	D	E
1, 2, 3 only	1, 3 only	2, 4 only	4 only	All are correct

A child experiencing an acute exacerbation of rheumatic fever usually has:

1. a prolonged PR interval
2. an increased antistreptolysin O titer
3. an increased erythrocyte sedimentation rate
4. subcutaneous nodules

R-TYPE ITEMS

Noted Flaws

Typically, items in an R-set are constructed around a common theme (eg, For each patient with abdominal pain, select the most likely diagnosis). For this example, the option set can contain an extensive list of causes of abdominal pain (up to 26 options) with several items (usually two to four) written to utilize that one option set. One of the main challenges with R-sets is the potential for multiple nonfunctional distractors (distractors that are not plausible enough for test-takers to choose). Often R-set authors create a lengthy list of options but do not take the opportunity to go back after they construct the items to ensure that there are at least three plausible distractors within the option set for each associated item. Statistical performance potentially suffers as a result. The format also risks increasing the reading load for the examinee given the number of options from which to choose.

Sample R-type item



DIRECTIONS: Each set of matching questions in this section consists of a list of lettered options (some of which may be in figures), followed by several numbered items. For each numbered item, select the ONE lettered option that is most closely associated with it. Each lettered option may be selected once, more than once, or not at all.

- (A) Articular cartilage
- (B) Endosteum
- (C) Epiphyseal plate
- (D) Epiphysis
- (E) Medullary cavity
- (F) Periosteal bone collar
- (G) Periosteum
- (H) Primary ossification center
- (I) Vascular osteogenic bud

For each description involving bone, select the most likely anatomic structure.

1. Mitoses in this structure are responsible for longitudinal growth of bone.
2. This structure contains a secondary center of ossification during bone development.

APPENDIX D: RESOURCES AND ADDITIONAL READING ON ITEM AND TEST DEVELOPMENT

- Case SM. Assessment of truths we hold as self-evident and their implications. In: Scherpbier AJJA, van der Vleuten CPM, Rethans JJ, van der Steeg AFW, eds. *Advances in Medical Education*. Dordrecht, The Netherlands: Kluwer Academic Publishers; 1997:2-6.
- Gronlund NE, Linn RL. *Measurement and Evaluation in Teaching*. 6th ed. New York, NY: Macmillan Publishers; 1990.
- Haladyna TM. *Writing Test Items to Evaluate Higher-Order Thinking*. Needham Heights, MA: Allyn & Bacon; 1997.
- Haladyna TM, Downing SM. A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*. 1989a;1:37-50.
- Haladyna TM, Downing SM. The validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*. 1989b;1:51-78.
- Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*. 2002;15(3):309-333.
- Haladyna TM, Rodriguez MC. *Developing and Validating Test Items*. New York, NY: Routledge; 2013.
- Henrysson S. Gathering, analyzing, and using data on test items. In: Thorndike RL, ed. *Educational Measurement*. Washington, DC: American Council on Education; 1971.
- Hubbard JP. *Measuring Medical Education*. Philadelphia, PA: Lea & Febiger; 1971.
- Hubbard JP, Clemens WV. *Multiple-Choice Examinations in Medicine*. Philadelphia, PA: Lea & Febiger; 1961.
- Kelley TL. The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology*. 1939;30:17-24.
- Linn RL, Gronlund NE. *Measurement and Assessment in Teaching*. 7th ed. New York, NY: Macmillan Publishers; 1995.
- Millman J, Greene J. The specification and development of tests of achievement and ability. In: Linn RL, ed. *Educational Measurement*. 3rd ed. Phoenix, AZ: Oryx Press; 1989:335-366.
- Newble DI, Dauphinee D, Woolliscroft JO, et al. Guidelines for assessing clinical competence. *Teaching and Learning in Medicine*. 1994;6(3):213-220.
- Norman G, Swanson DB, Case SM. Conceptual and methodological issues in studies comparing assessment formats. *Teaching and Learning in Medicine*. 1996;8(4):208-216.
- Osterlind SJ. *Constructing Test Items: Multiple-choice, Constructed-response, Performance, and Other Formats*. Norwell, MA: Kluwer Academic Publishers; 1998.
- Paniagua M, Swygert KA, Downing S. Writing high quality constructed-response and selected-response items. In: *Assessment in Health Professions Education, 2nd ed*. New York, NY: Routledge; 2019.
- Parshall CG, Harmes JC, Davey T, Pashley PJ. Innovative items for computerized testing. In: *Elements of Adaptive Testing*. New York, NY: Springer; 2009:215-230.

Rodriguez, MC. Selected-response item development. In: Lane S, Raymond MR, Haladyna TM, eds. *Test Development*. 2nd ed. New York, NY: Routledge; 2015:259–273.

Shen L, Li F, Wattleworth R, Filipetto F. The promise and challenge of including multimedia items in medical licensure examinations: some insights from an empirical trial. *Academic Medicine*. Oct 2010;85(10 suppl):S56–S59.

Swanson DB, Case SM. Assessment in basic science instruction: Directions for practice and research. *Advances in Health Sciences Education: Theory & Practice*. 1997;2:71–84.

