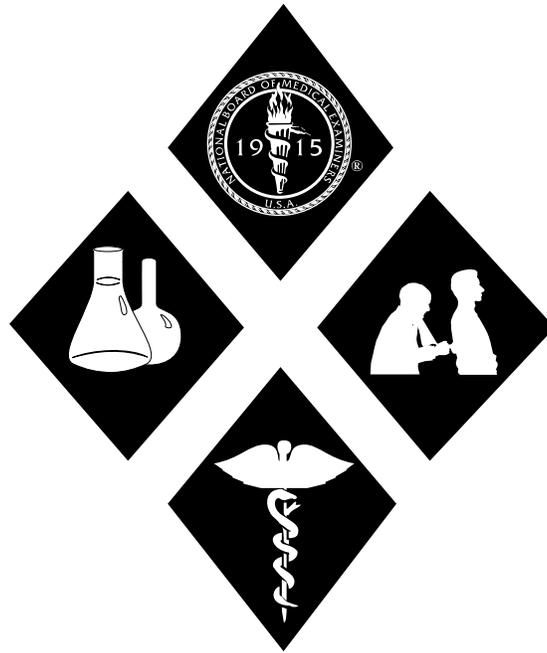

Constructing Written Test Questions For the Basic and Clinical Sciences

Third Edition
(Revised)



*National Board of Medical Examiners
3750 Market Street
Philadelphia, PA 19104*

Printed copies of this publication are not available from the National Board of Medical Examiners (NBME). Additional copies can be obtained by downloading the manual from the NBME's web site. Permission to copy and distribute this document is granted by the 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 the NBME.

Copyright © 1996, 1998 National Board of Medical Examiners® (NBME®).

Copyright © 2001, 2002 National Board of Medical Examiners® (NBME®).

All right reserved. Printed in the United States of America.

Constructing Written Test Questions For the Basic and Clinical Sciences

Third Edition
(Revised)

Contributing Authors

*Susan M. Case, PhD and David B. Swanson, PhD**

*National Board of Medical Examiners
3750 Market Street
Philadelphia, PA 19104*

**Dr. Case was a Senior Evaluation Officer at the National Board of Medical Examiners when the first three editions of this manual were published; she is currently the Director of Testing at the National Conference of Bar Examiners.*

Dr. Swanson is the Deputy Vice President of Professional Services at the National Board of Medical Examiners.

Table of Contents



	Page
Section I	
Issues Regarding Format and Structure of Test Questions	7
<i>Chapter 1. Introduction</i>	9
Assessment: An Important Component of Instruction	9
Issues of Sampling	10
Importance of Psychometric Considerations	11
<i>Chapter 2. Multiple-Choice-Item Formats</i>	13
True/False vs One-Best-Answer Questions	13
The True/False Family	14
The One-Best-Answer Family	16
The Bottom Line on Item Formats	18
<i>Chapter 3. Technical Item Flaws</i>	19
Issues Related to Testwiseness	19
Issues Related to Irrelevant Difficulty	22
Summary of Technical Item Flaws	26
Use of Imprecise Terms in Examination Questions	27
Section II	
Writing One-Best-Answer Questions for the Basic and Clinical Sciences	31
The Basic Rules for One-Best-Answer Items	33
<i>Chapter 4. Item Content: Testing Application of Basic Science Knowledge</i>	35
Item Content for the Basic Sciences	35
Item Templates	38
Additional Templates	39
Types of Questions and Sample Lead-ins and Option Lists	40
Writing the Options: Altering Item Difficulty	41
Item Shape	42
Problem-Based Learning and Use of Case Clusters	43
Sample Items for the Basic Sciences	47
<i>Chapter 5. Item Content: Testing Application of Clinical Science Knowledge</i>	51
Methods for Assessment	51
General Issues Regarding What to Test	52

Testing Recall of Isolated Facts or Application of Knowledge	53
Writing One-Best-Answer Items	56
Fine Points on Item Stems	57
Verbosity, Window Dressing, and Red Herrings:	
Do They Make a Better Test Item?	58
Writing Items Related to Physician Tasks	61
Writing Items on Difficult Topics	66
Section III	
Extended-Matching Items	69
<i>Chapter 6. Extended-Matching (R-Type) Items</i>	71
Avoiding Flaws When You Write Extended-Matching Items for Your Own Examination	72
Sample Lead-ins and Topics for Option Lists	74
More on Options for R-Sets	75
Writing the Item Stems	76
Sample Good and Bad Item Stems Using the Same Option List	77
Overview of the Steps for Writing Extended-Matching Items	81
Sample Extended-Matching Sets	82
Steps for Organizing a Group to Write Clinical R-sets	90
Form for Writing R-Sets	93
Sample SPSSX Code to Score Multiple-Choice Tests Including Extended-Matching Items	94
Comparison of Items in Five-Option and Extended-Matching Format	96
A's to R's and Back Again	97
<i>Chapter 7. Pick N Items: An Extension of the Extended-Matching Format</i>	99
Section IV	
Additional Issues	105
<i>Chapter 8. Interpretation of Item Analysis Results</i>	107
<i>Chapter 9. Establishing a Pass/Fail Standard</i>	111
Definitions and Basic Principles	111
Two Standard-Setting Methods Based on Judgements about Items	112
Relative/Absolute Compromise Standards: The Hofstee Method	114
<i>Chapter 10. Miscellaneous Thoughts on Topics Related to Testing</i>	115
<i>Appendix A. The Graveyard of NBME Item Formats</i>	117
<i>Appendix B. Sample Item-Writing Templates, Items, Lead-Ins, and Option Lists For the Basic and Clinical Sciences</i> ..	129

Preface to Third Edition



This manual was written to help faculty members improve the quality of the multiple-choice questions written for their examinations. The manual provides an overview of item formats, concentrating on the traditional one-best-answer and matching formats. It reviews issues related to technical item flaws and issues related to item content. The manual also provides basic information to help faculty review statistical indices of item quality after test administration. An overview of standard-setting techniques is also provided. Issues related to exam blueprinting are not addressed in any detail. We have focused almost exclusively on the item level, leaving exam level planning for another manuscript.

We anticipate that this manual will be useful primarily by faculty who are teaching medical students in basic science courses and clinical clerkships. The examples focus on undergraduate medical education, though the general approach to item writing may be useful for assessing examinees at other levels.

This manual reflects lessons that we have learned in developing items and tests over the past 20 years. During this period, we have reviewed (quite literally) tens of thousands of multiple-choice questions and have conducted item-writing workshops for thousands of item writers preparing USMLE, NBME, and specialty board examinations as well as faculty at more than 60 medical schools developing test questions for their own examinations. Each workshop attendee has helped us to frame our thoughts regarding how to write better quality test questions, and, over the years, we have become better able (we believe) to articulate the why's and wherefore's. We hope this manual helps to communicate these thoughts.

Susan M. Case, PhD
David B. Swanson, PhD

January 1998

Section I

Issues Regarding Format and Structure of Test Questions



This section reviews structural issues important for the construction of high-quality test questions. The following section will review issues related to item content.

Chapter 1

Introduction

❖

Assessment: An Important Component of Instruction

Assessment is a critical component of instruction; properly used, it can aid in accomplishing key curricular goals. The impact of decisions regarding how and when to evaluate the knowledge and performance of your students cannot be overestimated.

A primary purpose of testing is to communicate what you view as important. Tests are a powerful motivator, and students will learn what they believe you value. Assessment also helps to fill instructional gaps by encouraging students to read broadly on their own and to participate broadly as educational opportunities are available. This outcome of testing is especially important in the clerkships, where the curriculum may vary from student to student, depending on factors such as the clinical setting and the random flow of patients. This outcome may also be important in some basic-science settings (eg, problem-based learning), where the educational experiences may vary from student to student.

Because tests have such a powerful influence on student learning, it is important to develop tests that will further your educational goals. Introduction of a hands-on clinical skills test drives students out of the library into the clinic, where they may seek help with their physical-exam skills; introduction of a test assessing only recall of isolated facts, on the other hand, drives them to “cram” course review books. This manual focuses on how to write high-quality, multiple-choice questions that assess skill in interpreting data and making decisions, which we believe are important components of clinical skills.

Students’ paths toward mastery or even excellence will be less rocky if they receive ongoing feedback on their progress.

Purposes of Testing

- Communicate to students what material 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 where the course/curriculum is weak

What Should Be Tested?

- Exam content should match course/clerkship objectives
- Important topics should be weighted more heavily than less important topics
- The testing time devoted to each topic should reflect the relative importance of the topic
- The sample of items should be representative of the instructional goals

Issues of Sampling

The purpose of any assessment is to permit inferences to be drawn concerning the skills of examinees: inferences that extend beyond the particular problems (or, equivalently, cases or test questions) included in the exam to the larger domain from which the cases (or questions) are sampled. It's clear to all of us that assessment takes time. It's also clear that, if you increase time spent in one activity, you have to decrease time spent in other activities. Whether you're deciding on an overall plan for evaluation or you're deciding what to include on a single test, you're basically faced with a sampling problem. Performance on the sample provides a basis for estimating achievement in the broader domain that is actually of interest.

With multiple-choice questions (MCQs), you first need to decide what you want to include on the test. The amount of attention given to evaluating something should reflect its relative importance. You need to sample topics and also sample skills (eg, determining the diagnosis, deciding on the next step in management); you cannot ask everything. Performance on the sample provides a basis for estimating achievement in the broader domain that is actually of interest. The nature of the sample determines the extent to which the estimate of true ability is reproducible (reliable, generalizable) and accurate (valid). If the sample is not representative of the broader domain of interest (eg, including only cardiovascular-related content in a

test of competence in general medical practice), exam results will be biased and will not provide a good basis for estimating achievement in the domain of interest. If the sample is too small, exam results may not be sufficiently precise (reproducible, reliable) to ensure that they reflect true proficiency.

With a multiple-choice test, there's almost always one grader (usually the computer) and a series of questions or sets of questions; sampling involves selecting a subset of questions to include on the test. With other evaluation methods (eg, oral exams based on patient cases, standardized patient exams, essay exams), the sampling is much more complicated. Any method that can't be scored mechanically requires sampling on a second dimension: the dimension of grader. In these exams, you are interested in performance across a range of cases and you want the grade to be independent of who the examiner is. You therefore need to sample across two dimensions: one for the questions or cases and one for the judges or raters. You need to sample across a range of cases, because performance on one case is not a very good predictor of performance on other cases. You also need to sample across different raters to minimize the effects of rater harshness or leniency, and other issues like halo that cause problems in the consistency of scoring across raters. With broad samples, peaks and valleys in performance and peaks and valleys in rater differences tend to average out.

Although this manual focuses on multiple-choice questions, we believe that it is generally appropriate to use a variety of testing methods. No one method is likely to assess all the skills of interest. It should also be noted that the method used for assessment does not directly affect test quality, nor does it determine the component of competence measured by the test.

Importance of Psychometric Considerations

The extent to which the psychometric characteristics of an assessment method are important is determined by the purpose of the test and the decisions that will be made based on the results. For "high-stakes" tests (those used for promotion or graduation decisions, even course grades), test results must be reasonably reproducible (precise, reliable) and accurate (valid). For "low-stakes" tests, the psychometric characteristics are less important, and the primary consideration should be on directing student learning. As noted above, in order to generate a reproducible score, you need to sample content broadly (ie, typically, a dozen or more cases, 100 or more MCQs or short-answer items).

The following papers include more detail about assessment issues in general.

Swanson DB. (1987) A measurement framework for performance-based tests. In Hart I, Harden R (Eds.), *Further Developments in Assessing Clinical Competence*. Montreal: Can-Heal Publications, 13-4.

Swanson DB, Norcini JJ, Grosso L. (1987) Assessment of clinical competence: written and computer-based simulations. *Assessment and Evaluation in Higher Education*, 12(3), 220-246.

Friedman C, de Bliet R, Greer D, Mennin S, Norman G, Sheps C, Swanson DB, Woodward C. (1990) Charting the winds of change: recommendations for evaluating innovative medical curricula. *Academic Medicine*, 65, 8-14.

Swanson DB, Case SM, van der Vleuten CP. (1991:1997/2nd edition) Strategies for student assessment. In Boud, D. & Feletti, G (Eds.), *The Challenge of Problem-Based Learning*. London, Kogan Page Ltd., 269-282.

Newble DI, Dauphinee D, Woolliscroft JO, MacDonald M, Mulholland H, Page G, Swanson DB, Thomson A, van der Vleuten CP. (1994) Guidelines for assessing clinical competence. *Teaching and Learning in Medicine*, 6:3, 213-220.

Swanson DB, Norman G, Linn R. (1995) Performance-based assessment: Lessons from the health professions. *Educational Researcher*, 24:5, pp5-11,35. Abridged version reprinted in *Pedagogue: Perspectives on Health Sciences Education*, 6(Summer 1996), McMaster University Program for Educational Development, 1-7.

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

Case SM, (1997) 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, 2-6.

Swanson DB, Clauser BE, Case SM. (1999) Clinical skills assessment with standardized patients in high-stakes tests: A framework for thinking about score precision, equating, and security. *Advances in Health Sciences Education*, 4: 67-106.

The following papers discuss item format issues in greater detail.

Case SM, Downing SM. (1989) Performance of various multiple-choice item types on medical specialty examinations: types A, B, C, K, and X. *Proceedings of the 28th Annual Conference on Research in Medical Education*, 167-172.

Swanson DB, Case SM. (1992) Trends in written assessment: a strangely biased perspective. In Harden R, Hart I, Mulholland H (Eds.), *Approaches to the Assessment of Clinical Competence: Part 1*. Page Brothers, Norwich, England, 38-53.

Case SM, Swanson DB. (1993) Extended-matching items: a practical alternative to free-response questions. *Teaching and Learning in Medicine*, 5(2), 107-115.

Swanson DB, Case SM. (1995) Variation in item difficulty and discrimination by item format on Part I (basic sciences) and Part II (clinical sciences) of U.S. licensing examinations. In Rothman A, Cohen R (Eds.), *Proceedings of the Sixth Ottawa Conference on Medical Education*. University of Toronto Bookstore Custom Publishing, 285-287.

Norman G, Swanson DB, Case SM. (1996) Conceptual and methodological issues in studies comparing assessment formats. *Teaching and Learning in Medicine*, 8(4):208-216.

Chapter 2

Multiple-Choice-Item Formats

❖

In order for a test question to be a good one, it must satisfy two basic criteria. First, the test question must address important content. This is an essential condition, which will be addressed further along in the manual. Obviously, item content is of critical importance, but, in and of itself, focusing on important content is not sufficient to guarantee that your test question is a good one. Items that attempt to assess critically important topics cannot do so unless they are well-structured — avoiding flaws that benefit the testwise examinee and avoiding irrelevant difficulty are prerequisites that must be met in order for test questions to generate valid scores.

True/False vs One-Best-Answer Questions

The universe of multiple-choice questions (MCQ's) can be divided into two families of items: those that require the examinee to indicate all responses that are appropriate (true/false) and those that require the examinee to indicate a single response (one best answer).

Each family is represented by several specific formats, as listed below:

True/false-item formats require that examinees select all options that are true

- C (A / B/ Both/ Neither items)
- K (complex true/false items)
- X (simple true/false items)
- Simulations such as Patient Management Problems (PMPs)

One-best-answer item formats require that examinees select the single best response

- A (4 or more options, single items or sets)
- B (4 or 5 option matching items in sets of 2-5 items)
- R (Extended-Matching items in sets of 2-20 items)

The letters used to label the item formats hold no intrinsic meaning. Letters have been assigned more or less sequentially to new item formats as they are developed (see Appendix A).

The True/False Family

The true/false and one-best-answer families pose very different tasks for the examinee. True/false items require an examinee to select all the options that are “true.” For these items, the examinee must decide where to make the cut-off — to what extent must a response be “true” in order to be keyed as “true.” While this task requires additional judgement (beyond what is required in selecting the one best answer), this additional judgment may be unrelated to clinical expertise or knowledge. Too often, examinees have to guess what the item writer had in mind because the options are not either completely true or completely false.

The following is an example of an 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.

Which of the following is/are X-linked recessive conditions?

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

The options can be diagramed as follows.

2	1
4	3
<hr/>	
Totally Wrong Options	Totally Correct Options

True statements about cystic fibrosis (CF) include:

1. The incidence of CF is 1:2000.
2. Children with CF usually die in their teens.
3. Males with CF are sterile.
4. CF is an autosomal recessive disease.

This true/false item is flawed. Options 1, 2, and 3 cannot be judged as absolutely true or false; a group of experts would not agree on the answers. In thinking about Option 1, note that the incidence is not exactly 1:2000; experts would want more information: Is this in the USA? Is this among all ethnic groups? Modifying the language to “approximately 1:2000” doesn’t help, since the band is not specified. Similar issues arise with Options 2 and 3, while Option 4 is clear.

*Following tradition, for true/false items, the options are numbered; for one-best-answer items, the options are lettered.

While written in jest (by the second author), this true/false item illustrates a common problem — items for which the stem is unclear. Depending on your perspective, Options 1, 2, and 3 might be true; alternatively, 1, 2, and 3 might be false while 4 is true.

The way to a man's heart is through his

1. *aorta*
2. *pulmonary arteries*
3. *pulmonary veins*
4. *stomach*

In this true/false example, there are vague terms in the options that provide cues to the testwise examinee. For example, the term “may” in Options 1, 2, and 3 cues the testwise examinee that those options are true. Option 4 is harder to guess — what does “usually” mean? Research has shown that these vague frequency terms do not have a shared definition. Experts would not agree on whether the fourth option is true or false.

In the clinical assessment of chronic pain,

1. *the physician's personal attitude concerning pain may affect medical judgement*
2. *unpleasant emotions may be converted to complaints of bodily pain*
3. *pain may have a symbolic meaning*
4. *facial appearance or body posture is usually a clue to the severity of the pain*

The flaws in this item are more subtle. The difficulty is that the examinee has to make assumptions about the severity of the disease, the age of the patient, and whether or not the disease has been treated. Different assumptions lead to different answers, even among experts.

In children, ventricular septal defects are associated with

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

Note that in each sample flawed item, the stem is unclear, the options contain vague terms, or the options are partially correct. In each instance, a group of experts would have difficulty reaching a consensus on the correct answer.

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

- Stems must be clear and unambiguous. Imprecise phrases such as *is associated with*; *is useful for*; *is important* and words that provide cueing such as *may* or *could be*; and vague terms such as *usually* or *frequently* should be avoided.
- Options must be absolutely true or false; no shades of gray are permissible; avoid phrases and words noted in the first item above.

The One-Best-Answer Family

In contrast to true/false questions, one-best-answer (A-type) questions make explicit the number of options to be selected. A-type items are the most widely used multiple-choice-item format. They consist of a stem (eg, a clinical case presentation) and a lead-in question, followed by a series of choices, typically one correct answer and four distractors. The following question describes a situation (in this instance, a patient) and asks the examinee to indicate the most likely cause of the problem.

Stem:

A 32-year-old man has 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 F), blood pressure is 130/80 mm Hg, pulse is 94/min, and respirations are 42/min and shallow. 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; the plantar responses are flexor.

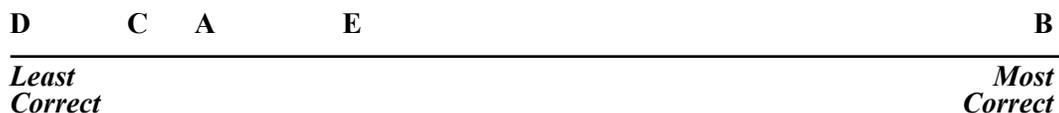
Lead-in:

Which of the following is the most likely diagnosis?

Options:

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

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

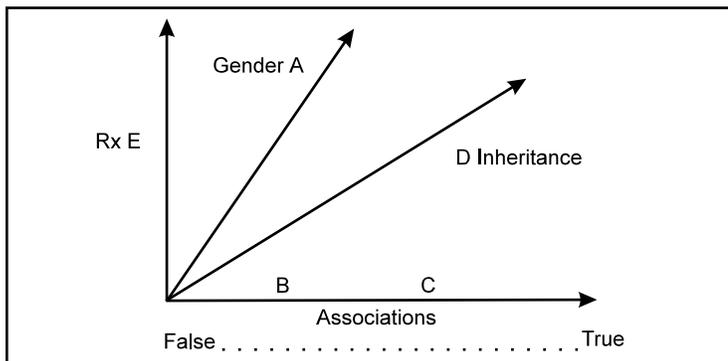


Even though the incorrect answers are not completely wrong, they are less correct than the “keyed answer.” The examinee is instructed to select the “most likely diagnosis”; experts would all agree that the most likely diagnosis is B; they would also agree that the other diagnoses are somewhat likely, but less likely than B. As long as the options can be laid out on a single continuum, in this case from “Most Likely Diagnosis” to “Least Likely Diagnosis,” options in one-best-answer questions do not have to be totally wrong.

This item is flawed. After reading the stem, the examinee has only the vaguest idea what the question is about. In an attempt to determine the “best” answer, the examinees have to decide whether “it occurs frequently in women” is more or less true than “it is seldom associated with acute pain in a joint.” This is a comparison of apples and oranges. In order to rank-order the relative correctness of options, the options must differ on a single dimension or else all options must be absolutely 100% true or false.

Which of the following is true about pseudogout?

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



The diagram of these options might look like this. 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 question appears to assess knowledge of several different points, its inherent flaws preclude this. The question by itself is not clear; the item cannot be answered without looking at the options.

In contrast to the options in the item on pseudogout, the options in the item on Guillain-Barré syndrome are homogeneous (eg, all diagnoses); knowledgeable examinees can rank-order the options along a single dimension.

Well-constructed one-best-answer questions satisfy the “cover-the-options” rule. The questions could be administered as write-in questions. The entire question is included in the stem.

The Bottom Line on Item Formats

We recommend that you do not use true/false questions. While many item writers believe the true/false items are easier to write than one-best-answer items, we find that they are more problematic. The item writer had something particular in mind when the question was written, but careful review commonly 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 rewrite or discard true/false items far more frequently than items written in other formats. Some ambiguities can be clarified, but others cannot.

There is a final reason that is more compelling than those noted above. We find that, to avoid ambiguity, we are pushed toward assessing recall of an isolated fact — something we are actively trying to avoid. We find that application of knowledge, integration, synthesis, and judgement questions can better be assessed by one-best-answer questions. As a result, the NBME has completely stopped using true/false formats on its examinations.

We also recommend that you not use negative A-type questions. The most problematic are those that take the form: “Each of the following is correct EXCEPT” or “Which of the following statements is NOT correct?” These suffer from the same problem as true/false questions: if options cannot be rank-ordered on a single continuum, the examinees cannot determine either the “least” or the “most” correct answer. On the other hand, we occasionally use well-focused negative A-types with single-word options on some exams, largely as a (poor) substitute for items that instruct the examinee to select more than one response. A superior format for this purpose, the Pick “N” format, in which examinees are instructed to select “N” responses, is discussed later in the manual.

The Appendix A illustrates a variety of item formats that are no longer used on NBME exams.

Chapter 3

Technical Item Flaws



This section describes two types of technical item flaws: testwiseness and irrelevant difficulty. Flaws related to testwiseness make it easier for some students to answer the question correctly, based on their test-taking skills alone. These flaws commonly occur in items that are unfocused and do not satisfy the “cover-the-options” rule. Flaws related to irrelevant difficulty make the question difficult for reasons unrelated to the trait that is the focus of assessment.

The purpose of this section is to outline common flaws and to encourage you to eliminate these flaws from your questions to provide a level playing field for the testwise and not-so-testwise students. The probability of answering a question correctly should relate to the examinee’s amount of expertise on the topic being assessed and should not relate to their expertise on test-taking strategies.

Issues Related to Testwiseness

Grammatical cues: one or more distractors don’t follow grammatically from the stem

Because an item writer tends to pay more attention to the correct answer than to the distractors, grammatical errors are more likely to occur in the distractors. In this example, testwise students would eliminate A and C as options because they do not follow grammatically or logically from the stem. Testwise students then have to choose only between B, D, and E.

A 60-year-old man is brought to the emergency department by the police, who found him lying unconscious on the sidewalk. After ascertaining that the airway is open, the first step in management should be intravenous administration of

- A. examination of cerebrospinal fluid*
- B. glucose with vitamin B₁ (thiamine)*
- C. CT scan of the head*
- D. phenytoin*
- E. diazepam*

Logical cues: a subset of the options are collectively exhaustive

In this item, Options A, B, and C include all possibilities. The testwise student knows that A, B, or C must be correct, whereas the non-testwise student spends time considering D and E. Often, the item writers add D and E only because they want to list five options. In these situations, the item writer may not have paid much attention to the merits of options D and E; sometimes, they are partially correct and confusing because they cannot be rank-ordered on the same dimension as Options A, B, and C. This flaw is commonly seen in items with options such as “Increases,” “Decreases,” and “Remains the same.”

Crime is

- A. equally distributed among the social classes*
- B. overrepresented among the poor*
- C. overrepresented among the middle class and rich*
- D. primarily an indication of psychosexual maladjustment*
- E. reaching a plateau of tolerability for the nation*

Absolute terms: terms such as “always” or “never” are used in options

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. Note that this flaw would not arise if the stem was focused and the options were short; it arises only when verbs are included in the options rather than in the lead-in.

In patients with advanced dementia, Alzheimer’s 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*

Long correct answer: correct answer is longer, more specific, or more complete than other options

In this item, Option C is longer than the other options; it is also the only double option. Item writers tend to pay more attention to the correct answer than to the distractors. Because you are teachers, you write long correct answers that include additional instructional material, parenthetical information, caveats, etc. Sometimes this can be quite extreme: the correct answer is a paragraph in length and the distractors are single words.

Secondary gain is

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

Word repeats: a word or phrase is included in the stem and in the correct answer

This item uses the word “unreal” in the stem, and “derealization” is the correct answer. Sometimes, a word is repeated only in a metaphorical sense, eg, a stem mentioning bone pain, with the correct answer beginning with the prefix “osteo-”.

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. This symptom is called

- A. depersonalization*
- B. derailment*
- C. derealization*
- D. focal memory deficit*
- E. signal anxiety*

Convergence strategy: the correct answer includes the most elements in common with the other options

This item flaw is less obvious than the others, but it occurs frequently and is worth noting. The flaw is seen in several forms. The underlying premise is that the correct answer is the option that has the most in common with the other options; it is not likely to be an outlier. For example, in numeric options, the correct answer is more often the middle number than an extreme value. In double options, the correct answer is more likely to be the option that has the most elements in common with the other distractors. For example, if the options are “Pencil and pen”; “Pencil and highlighter”; “Pencil and crayon”; “Pen and marker,” the correct answer is likely to be “Pencil and pen” (ie, by simple count, “Pencil” appeared 3 times in the options; “Pen” appeared twice; other elements each appeared only once). While this might seem ridiculous, this flaw occurs because item writers start with the correct answer and write permutations of the correct answer as the distractors. The correct answer is, therefore, more likely to have elements in common with the rest of the options; the incorrect answers are more likely to be outliers as the item writer has difficulty generating viable distractors. In this example, the testwise student would eliminate “anionic form” as unlikely because “anionic form” appears only once; that student would also exclude “outside the nerve membrane” because “outside” appears less frequently than “inside”. The student would then have to decide between Options B and D. Since three of the five options involve a charge, the testwise student would then pick Option B.

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*

Issues Related to Irrelevant Difficulty

Options are long, complicated, or double

This item illustrates a common flaw. The stem contains extraneous reading, but, more importantly, the options are very long and complicated. Trying to decide among these options requires a significant amount of reading because of the number of elements in each option. This can shift what is measured by an item from content knowledge to reading speed. Please note that this flaw relates only to options. There are many well-constructed test questions that include a long stem. Decisions about stem length should be made in accord with the purpose 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, then it is appropriate for the stem to include a fairly complete description of the situation.

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 assure 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, council, a chance to hear and confront evidence against him/her.*
- B. Proper notice, a tribunal empowered to make the decision, a chance to confront witnesses against him/her, and a chance to present evidence in defense.*
- C. Reasonable and timely notice, impartial panel empowered to make a decision, a chance to hear evidence against himself/herself and to confront witnesses, and the ability to present evidence in defense.*

Numeric data are not stated consistently

When numeric options are used, the options should be listed in numeric order and the options should be listed in a single format (ie, as single terms or as ranges). Confusion occurs when formats are mixed and when the options are listed in an illogical order or in an inconsistent format.

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.

Following a second episode of infection, what is the likelihood that a woman is infertile?

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

Frequency terms in the options are vague (eg, rarely, usually)

Research has shown that vague frequency terms are not consistently defined or interpreted, even by experts. A more complete discussion of this research is included on page 29.

Severe obesity in early adolescence

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

Language in the options is not parallel; options are in an illogical order

This item illustrates a common flaw in which the options are long and the language makes it difficult and time-consuming to determine which is the most correct. Generally, this flaw can be corrected by careful editing. In this particular item, the lead-in can be changed to “For which of the following reasons can no conclusion be drawn from these results?” The options can then be edited (ie, A. No follow-up was made of nonvaccinated children; B. The number of cases was too small; C. The trial involved only boys, and a new option can be written for D).

In a vaccine trial, 200 2-year-old boys were given a vaccine against a certain disease and then monitored for five 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. *No conclusion can be drawn, since no follow-up was made of nonvaccinated children*
- B. *The number of cases (ie, 30 cases over five years) is too small for statistically meaningful conclusions*
- C. *No conclusions can be drawn because the trial involved only boys*
- D. *Vaccine efficacy (%) is calculated as $85-15/100$*

“None of the above” is used as an option

The phrase “None of the above” is problematic in items where judgement is involved and where the options are not absolutely true or false. If the correct response is intended to be one of the other listed options, knowledgeable students can be faced with a dilemma because they have to decide between a very detailed perfect option and the one that you have intended as correct. They can often construct an option that is more correct than the one you intended to be correct. 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 will often be possible to fix such items by replacing “none of the above”

by an option that means roughly the same thing but is more specific. For example, in an item asking an examinee to specify the most appropriate pharmacotherapy, replacing “none of the above” by “no drug should be given at this time” will eliminate the ambiguity of “none of the above.”

Stems are tricky or unnecessarily complicated

Sometimes, item writers can take a perfectly easy question and turn it into something so convoluted that only the most stalwart will even read it. This item is a sample of that kind of item. The notation in I: through V: is complex; having to rank order Roman numerals after working through that notation is irrelevant and unnecessarily difficult.

Which city is closest to New York City?

- A. Boston
- B. Chicago
- C. Dallas
- D. Los Angeles
- E. none of the above

If students select E, you don't know if they are thinking about Philadelphia or London.

Arrange the parents of the following children with Down's 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

Summary of Technical Item Flaws

Issues Related to Testwiseness

- **Grammatical cues** - one or more distractors don't follow grammatically from the stem
- **Logical cues** - a subset of the options is collectively exhaustive
- **Absolute terms** - terms such as "always" or "never" are in some options
- **Long correct answer** - correct answer is longer, more specific, or more complete than other options
- **Word repeats** - a word or phrase is included in the stem and in the correct answer
- **Convergence strategy** - the correct answer includes the most elements in common with the other options

Issues Related to Irrelevant Difficulty

- Options are long, complicated, or double
- Numeric data are not stated consistently
- Terms in the options are vague (eg, "rarely," "usually")
- Language in the options is not parallel
- Options are in a nonlogical order
- "None of the above" is used as an option
- Stems are tricky or unnecessarily complicated
- The answer to an item is "hinged" to the answer of a related item

General Guidelines for Item Construction

- Make sure the item can be answered without looking at the options OR that the options are 100% true or false.
- Include as much of the item as possible in the stem; the stems should be long and the options short.
- Avoid superfluous information.
- Avoid "tricky" and overly complex items.
- Write options that are grammatically consistent and logically compatible with the stem; list them in logical or alphabetical order. Write distractors that are plausible and the same relative length as the answer.
- Avoid using absolutes such as *always*, *never*, and *all* in the options; also avoid using vague terms such as *usually* and *frequently*.
- Avoid negatively phrased items (eg, those with *except* or *not* in the lead-in). If you must use a negative stem, use only short (preferably single word) options.

And most important of all: Focus on important concepts; don't waste time testing trivial facts.

Use of Imprecise Terms in Examination Questions

While imprecise terms are used in our everyday speech and in our writing, these terms cause confusion when they are used in the text of examination items. In a study conducted at the NBME, 60 members of eight test committees who wrote questions for various medical specialty examinations reviewed a list of terms used in MCQs to express some concept related to frequency of occurrence and indicated the percentage of time that was reflected by each term.

Results (shown below) indicated that the terms do not have an operational definition that is commonly shared, even among the item writers themselves. The mean value plus or minus one standard deviation exceeded 50 percentage points for more than half of the phrases. For example, on average, the item writers believed the term *frequently* indicated 70% of the time; half believed it was between 45% and 75% of the time; actual responses ranged from 20% to 80%. Of particular note is that values for *frequently* overlapped with values for *rarely*.

The implication of these results for the construction of test questions varies by item format. Vague terms create far more severe problems in the various kinds of true/false items (K-, C- and X-type items) than in one-best-answer (A- and R-type) items. For example, imprecise terms cause major problems in true/false items such as this example:

True statements about pseudogout include:

- 1. It occurs commonly in women.*
- 2. It is often associated with acute pain.*
- 3. It is usually hereditary.*
- 4. Serum calcium levels are frequently increased.*

In true/false items, the examinee has to judge whether each option is true or false. When options are not absolutely true or false, examinees rely on their personal definition of the ambiguous terms or their guesses about what these terms meant to the item writer. Alternatively, examinee responses may reflect personal response style (the tendency to respond either true or false when the correct answer is unknown). These response style factors may have more of an effect on whether or not an examinee answers the item correctly than knowledge of the subject matter and may be part of the reason why true/false items tend to perform poorly.

Rewording the options by specifying exact numbers does not correct the problem. For example, the statement, “the incidence among women is 1:2000” would not be an appropriate modification of Option 1 in the example shown. The incidence is not exactly 1:2000, and because a band is not specified, examinees would define their own bands, narrowly or widely, presumably depending on personal response styles. In true/false items, the appropriate treatment of numeric options is either to generate a comparison (eg, the incidence is greater than that of osteoarthritis) or to specify a range (eg, the incidence is between 1:1000 and 1:2000).

The issue noted above with true/false items is not as problematic with well-constructed one-best-answer items (ie, those that pose a clear question and have homogeneous options). For example, the following question includes a vague term in the item stem, yet, because the task is to select the one-best answer, the question is relatively unambiguous.

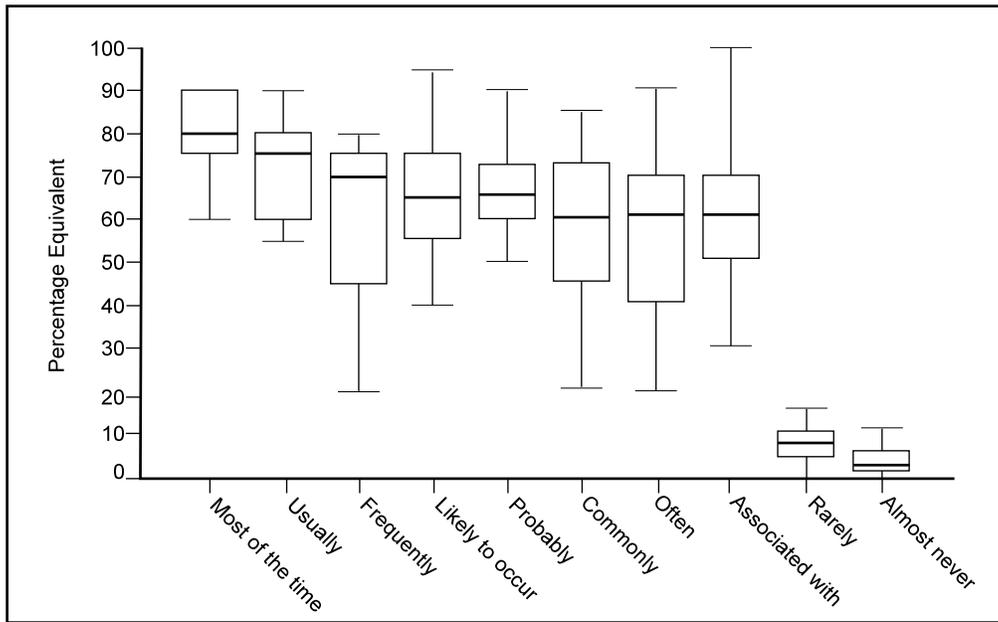
Which of the following laboratory values is usually increased in patients with pseudogout?

Problems do arise with one-best answer items that have vague terms in the options as in this example.

The only way to make such an item more ambiguous would be to use a fifth option “none of the above.”

Patients with pseudogout have pain:

- A. frequently*
- B. usually*
- C. often*
- D. commonly*



Box-plot showing distribution of responses for frequency terms. Results are based on responses from 60 members of eight item-writing committees. The horizontal line in each box indicates the median response; the boxes include the ranges for 50% of the responses. The vertical lines extend to the highest and lowest values indicated. For example, the median response for “frequently” indicated 70% of the time; half believed it was between 45% and 75% of the time; actual responses ranged from 20% to 80%, almost overlapping with “rarely.”

From: Case SM. (1994) The use of imprecise terms in examination questions: How frequent is frequently? *Academic Medicine*, 69(suppl):S4-S6.

Section II

Writing One-Best-Answer Questions for the Basic and Clinical Sciences



The previous chapters outlined technical issues related to the construction of multiple-choice questions. Section II focuses on item content.

The Basic Rules for One-Best-Answer Items

- ***Each item should focus on an important concept, typically a common or potentially catastrophic clinical problem.*** Don't waste testing time with questions assessing knowledge of trivial facts. Focus on problems that would be encountered in real life. Avoid trivial, "tricky," or overly complex questions.
- ***Each item should assess application of knowledge, not recall of an isolated fact.*** The item stems may be relatively long; the options should be short. Clinical vignettes provide a good basis for a question. For the clinical sciences, each should begin with the presenting problem of a patient, followed by the history (including duration of signs and symptoms), physical findings, results of diagnostic studies, initial treatment, subsequent findings, etc. Vignettes may include only a subset of this information, but the information should be provided in this specified order. For the basic sciences, patient vignettes may be very brief; "laboratory vignettes" are also appropriate.
- ***The stem of the item must pose a clear question, and it should be possible to arrive at an answer with the options covered.*** To determine if the question is focused, cover up the options and see if the question is clear and if the examinees can pose an answer based only on the stem. Rewrite the stem and/or options if they could not.
- ***All distractors (ie, incorrect options) should be homogeneous.*** They should fall into the same category as the correct answer (eg, all diagnoses, tests, treatments, prognoses, disposition alternatives). Rewrite any dissimilar distractors. Avoid using "double options" (eg, do W and X; do Y because of Z) unless the correct answer and all distractors are double options. Rewrite double options to focus on a single point. All distractors should be plausible, grammatically consistent, logically compatible, and of the same (relative) length as the correct answer. Order the options in logical order (eg, numeric), or in alphabetical order.
- ***Avoid technical item flaws that provide special benefit to testwise examinees or that pose irrelevant difficulty.***

Do **NOT** write any questions of the form "Which of the following statements is correct?" or "Each of the following statements is correct EXCEPT." These questions are unfocused and have heterogeneous options.

Subject each question to the five "tests" implied by the above rules. If a question passes all five, it is probably well-phrased and focused on an appropriate topic.

See also:

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

Chapter 4

Item Content:

Testing Application of Basic Science Knowledge



Item Content for the Basic Sciences

Traditionally, items are classified by the cognitive processes required to answer the question (eg, recall, interpretation, or problem solving; memory, comprehension, or reasoning). Recall items are thought to test examinees' knowledge of isolated facts. Interpretation items require examinees to review some information (often in tabular or graphical form) and reach some conclusion (eg, a diagnosis). Problem-solving items present a situation and require examinees to take some action (eg, the next step in patient management). Interpretation and problem-solving items are thought to involve "higher order" skills, rather than just rote memory of factual information.

Unfortunately, the cognitive processes required to answer an item are often difficult to determine, because they are as dependent on the background of the examinee as they are on the item content. For example, an item concerning blood flow in a patient with ventricular septal defect might require simply recall with little or no conscious thought from a pediatric cardiologist or cardiovascular physiologist, but a typical Step 1 examinee might have to reason out the answer from basic principles of hemodynamics. The cognitive processes involved in responding to a question vary by examinee, making this taxonomic approach difficult to use.

A simpler, more objective approach bases item classification on the *task* of the examinee. If an item requires an examinee to reach a conclusion, make a prediction, or select a course of action, it should be classified as an *application of knowledge* item. If an item tests only rote memory for isolated facts (without requiring their application), it should be classified as a *recall* item. *All items should require application of knowledge*, allowing assessment of both an examinee's information base plus ability to use that information.

The following pair of item stems illustrate the difference between a question assessing recall of an isolated fact and a question assessing application of knowledge.

Basic Science Recall Item Stem:

What area is supplied with blood by the posterior inferior cerebellar artery?

Basic Science Application of Knowledge Item Stem:

A 62-year-old man develops left-sided limb ataxia, Horner's syndrome, nystagmus, and loss of appreciation of facial pain and temperature sensations. What artery is most likely to be occluded?

It is common to use clinical vignettes as item stems to assess application of basic science knowledge to interpret clinical situations. For example, instead of asking examinees to identify the muscles innervated by a cranial nerve, provide a set of physical findings and ask examinees to identify the most likely site of the lesion. Instead of asking for a description of respiratory acidosis or alkalosis, provide values for arterial blood gases (and other patient findings as needed) and ask examinees to identify the most likely pathophysiologic explanation. Make sure that examinees can answer the question based on an understanding of basic science; experience in patient care should not be necessary.

“Lab vignettes” can also be useful in preparing items that test application of knowledge. These items present lab experiments and require examinees to use their understanding of basic science principles to predict or explain the results. The vignettes may describe classic experiments in a basic science area, or they may involve less well-known or hypothetical situations. Such items effectively shift the focus of assessment from knowledge of isolated facts to use of basic science principles to solve problems.

Use of patient and lab vignettes to assess application of knowledge has several benefits. First, the “face validity” of the exam is greatly enhanced by using “problem-solving” items. Second, items are more likely to focus on important information, rather than trivia. Third, it helps to identify those examinees who have memorized a substantial body of factual information, but are unable to use that information effectively.

Guidelines for Basic Science Item Content

- Test application of knowledge using experimental and clinical vignettes
- Focus items on key concepts and principles that are essential information (without access to references) for all examinees to understand
- Test material that is relevant to learning in clinical clerkships, postgraduate medical education, and beyond
- Avoid items that *only* require recall of isolated facts
- Avoid esoteric or interesting topics that are not essential

These two items were written to assess the same topic. We recommend that questions be written like the second item, not the first one.

Acute intermittent porphyria is the result of a defect in the biosynthetic pathway for

- A. collagen
- B. corticosteroid
- C. fatty acid
- D. glucose
- *E. heme
- F. thyroxine (T_4)

An otherwise healthy 33-year-old man has mild weakness and occasional episodes of steady, severe abdominal pain with some cramping but no diarrhea. One aunt and a cousin have had similar episodes. During an episode, his abdomen is distended, and bowel sounds are decreased. Neurologic examination shows mild weakness in the upper arms. These findings suggest a defect in the biosynthetic pathway for

- A. collagen
- B. corticosteroid
- C. fatty acid
- D. glucose
- *E. heme
- F. thyroxine (T_4)

Item Templates

The overall structure of an item can be depicted by an item template. You can typically generate many items using the same template. For example, the following template could be used to generate a series of questions related to gross anatomy:

A (patient description) is unable to (functional disability). Which of the following is most likely to have been injured?

This is a question that could be written using this template:

A 65-year-old man has difficulty rising from a seated position and straightening his trunk, but he has no difficulty flexing his leg. Which of the following muscles is most likely to have been injured?

- *A. Gluteus maximus*
- B. Gluteus minimus*
- C. Hamstrings*
- D. Iliopsoas*
- E. Obturator internus*

Many basic science questions can be presented within the context of a patient vignette. The patient vignettes may include some or all of the following components:

Age, Gender (eg, A 45-year-old man)

Site of Care (eg, comes to the emergency department)

Presenting Complaint (eg, because of a headache)

Duration (eg, that has continued for 2 days).

Patient History (with Family History ?)

Physical Findings

+/- Results of Diagnostic Studies

+/- Initial Treatment, Subsequent Findings, etc.

Additional Templates

A (*patient description*) has a (*type of injury and location*). Which of the following structures is most likely to be affected?

A (*patient description*) has (*history findings*) and is taking (*medications*). Which of the following medications is the most likely cause of his (*one history, PE or lab finding*)?

A (*patient description*) has (*abnormal findings*). Which [additional] finding would suggest/suggests a diagnosis of (*disease 1*) rather than (*disease 2*)?

A (*patient description*) has (*symptoms and signs*). These observations suggest that the disease is a result of the (*absence or presence*) of which of the following (*enzymes, mechanisms*)?

A (*patient description*) follows a (*specific dietary regime*). Which of the following conditions is most likely to occur?

A (*patient description*) has (*symptoms, signs, or specific disease*) and is being treated with (*drug or drug class*). The drug acts by inhibiting which of the following (*functions, processes*)?

A (*patient description*) has (*abnormal findings*). Which of the following (*positive laboratory results*) would be expected?

(*time period*) after a (*event such as trip or meal with certain foods*), a (*patient or group description*) became ill with (*symptoms and signs*). Which of the following (*organisms, agents*) is most likely to be found on analysis of (*food*)?

Following (*procedure*), a (*patient description*) develops (*symptoms and signs*). Laboratory findings show (*findings*). Which of the following is the most likely cause?

A (*patient description*) dies of (*disease*). Which of the following is the most likely finding on autopsy?

A patient has (*symptoms and signs*). Which of the following is the most likely explanation for the (*findings*)?

A (*patient description*) has (*symptoms and signs*). Exposure to which of the (*toxic agents*) is the most likely cause?

Which of the following is the most likely mechanism of the therapeutic effect of this (*drug class*) in patients with (*disease*)?

A patient has (*abnormal findings*), but (*normal findings*). Which of the following is the most likely diagnosis?

See Appendix B for additional examples.

Types of Questions

Guess my drug
Guess my toxic exposure
Guess my diet
Guess my mood

Predict physical findings
Predict lab findings
Predict sequelae

Identify underlying cause/diagnosis
Identify cause of drug responses
Identify drug to administer

Sample Lead-ins and Option Lists

Which of the following is (abnormal)?

Options sets could include sites of lesions; list of nerves; list of muscles; list of enzymes; list of hormones; types of cells; list of neurotransmitters; list of toxins, molecules, vessels, spinal segments.

Which of the following findings is most likely?

Options sets could include list of laboratory results; list of additional physical signs; autopsy results; results of microscopic examination of fluids, muscle or joint tissue; DNA analysis results; serum levels.

Which of the following is the most likely cause?

Options sets could include list of underlying mechanisms of the disease; medications that might cause side effects; drugs or drug classes; toxic agents; hemodynamic mechanisms, viruses, metabolic defects.

Which of the following should be administered?

Options sets could include drugs, vitamins, amino acids, enzymes, hormones.

Which of the following is defective/deficient/nonfunctioning?

Options sets could include list of enzymes, feedback mechanisms, endocrine structures, dietary elements, vitamins.

Given the pedigree, what is the likelihood that the next child (specify gender) will have the disease?

Writing the Options: Altering Item Difficulty

The incorrect options in each question are called distractors. Each distractor should be selected by some examinees; therefore, each distractor should be plausible and none should stand out as being obviously incorrect. Common misconceptions and faulty reasoning provide a good source of plausible distractors. Distractors directly affect the difficulty of a question. Consider the question to the right.

Who was the primary author of the Declaration of Independence?

- A. *Abraham Lincoln*
- B. *Thomas Jefferson*
- C. *Franklin Roosevelt*
- D. *King George II*
- E. *Catherine the Great*

In the example above, the options are quite divergent and Thomas Jefferson is easily identified as the correct answer. Someone who knows relatively little about American history could answer this correctly.

Now consider the same question with a different set of options.

In this example, the question becomes more difficult; the options are all plausible answers to someone who has limited knowledge. For some content areas, options like those in the first example might be appropriate; for others, those in the second example are more appropriate.

Who was the primary author of the Declaration of Independence?

- A. *George Washington*
- B. *Thomas Jefferson*
- C. *Alexander Hamilton*
- D. *Benjamin Franklin*
- E. *James Madison*

When writing your options, make sure that they are:

- Homogeneous in content (eg, all are diagnoses; all are next steps in patient care)
- Incorrect or inferior to the correct answer
- Plausible and attractive to the uninformed
- Similar to the correct answer in construction and length
- Grammatically consistent and logically compatible with the stem

Item Shape

An appropriately shaped item includes as much of the item as possible in the stem; the stem should be relatively long and the options should be relatively short. The stem should include all relevant facts; no additional data should be provided in the options.

Appropriately Shaped Item:

Long Stem

- A.
- B.
- C. Short Options
- D.
- E.

Poorly Shaped Item:

Short Stem

- A.
- B.
- C. Long Options
- D.
- E.

Problem-Based Learning and Use of Case Clusters

An increasing number of medical schools have adopted problem-based learning (PBL) as an instructional strategy for portions of the basic science curriculum. Although each school's approach to PBL is somewhat unique, all involve the use of written patient cases (problems) in basic science instruction. Problems are designed to stimulate learning of material from traditional basic science disciplines (eg, anatomy, physiology, biochemistry) from a clinical perspective, and application of basic science principles to clinical situations is stressed. Material is typically covered through independent study and discussed in small groups with a faculty tutor.

PBL courses and curricula typically emphasize the learning process, learning how to learn, responsibility of students for their own learning, and preparation for lifelong learning. However, there are important variations among programs that have implications for assessment. The Open Discovery approach emphasizes the learning process: students have responsibility for determining what to learn, as well as when and how to learn it. Learning to apply broad principles in problem-solving situations is viewed as most important, with minimal guidance provided by instructors and maximum opportunity for exploration by students. In contrast, in the Guided Discovery approach, curriculum developers identify specific learning objectives for each problem, and these objectives are provided to instructors who use them to organize group discussion and student learning. These curricula can be highly structured, with careful sequencing of instructional experiences. Students may or may not be aware of the structure and the specific objectives: their experience may be quite similar to students in programs using the Open Discovery approach. In practice, the Open and Guided Discovery approaches are probably best viewed as opposite ends of a continuum. Programs vary along the continuum, and, within a program, problems (and groups) also vary.

Assessment in programs using the Open Discovery approach often focuses on process variables such as self-directedness, motivation, effort, problem-solving, and attitudes. Assessment of learning outcomes is genuinely problematic, because each student is encouraged to pursue a somewhat different course of study. Use of traditional multiple-choice tests, in particular, is often viewed as inappropriate, because they may cause students to "study to the test," thus discouraging students from self-determination of the material to be learned and the process for learning it.

Assessment of learning outcomes poses fewer problems when the Guided Discovery approach is used, since the same learning objectives that guide problem development and use can also guide test development. To achieve congruence with curricular goals, assessment should focus on students' understanding of basic mechanisms of health, disease and treatment. Well-written multiple-choice tests can play a major role in assessment, as long as they stress application of basic science

knowledge to patient care. Tests using “case clusters” — multiple-choice questions associated with the same patient presentation — are particularly appropriate for PBL courses.

An example of a simple case cluster is shown below. It consists of a brief case presentation, followed by a series of three multiple choice questions. Each question addresses a somewhat different aspect of the case, looking at the clinical situation from a variety of perspectives. Like PBL more generally, use of test material like this emphasizes learning of basic science information so that it is organized to be useful in provision of patient care.

A 34-year-old woman has had severe watery diarrhea for the past four days. Two months earlier she had infectious mononucleosis. She abuses drugs intravenously and has antibodies to HIV in her blood. Physical examination shows dehydration and marked muscle weakness.

1. Laboratory studies are most likely to show
 - A. decreased serum K^+ concentration
 - B. decreased serum Ca^{2+} concentration
 - C. increased serum HCO_3^- concentration
 - *D. increased serum Na^+ concentration
 - E. increased serum pH
2. In evaluating the cause of the diarrhea, which of the following is most appropriate?
 - A. Colonic biopsy to identify *Giardia lamblia*
 - B. Culture of the oral cavity for *Candida albicans*
 - C. Duodenal biopsy to identify *Entamoeba histolytica*
 - D. Gastric aspirate to identify *Mycobacterium avium-intracellulare*
 - *E. Stool specimen to identify *Cryptosporidium*
3. Further studies to evaluate her HIV infection show the ratio of helper T lymphocytes to suppressor T lymphocytes to be 0.3. This occurs because HIV
 - A. induces proliferation of helper T lymphocytes
 - B. induces proliferation of suppressor T lymphocytes
 - *C. infects cells with CD4 receptors
 - D. infects macrophages
 - E. stimulates the synthesis of leukotriene

In addition to principles described earlier in this manual, there are two more considerations required in preparing case clusters: cueing and hinging. First, it is desirable to avoid “cueing” — providing hints at the answers to earlier questions in later questions. Students are very likely to “read ahead” for these clues, and item writers should avoid providing them. For example, in a cluster describing a patient with chest pain, if the first question addresses the most likely cause of the pain and the second requires selection of the most appropriate drug treatment, it is important that each of the diagnoses associated with the first question have a “matching” drug in the second (and vice versa); testwise examinees can rule out diagnoses (and drugs) simply by comparing the option lists.

Second, it is desirable to avoid “hinging” — creating questions where students must know the answer to one question in order to answer other questions — unless the topic to be tested is so important that the item writer is willing to have students receive either all of the points or none of the points associated with a cluster. The cluster on the next page, prepared by Drs. David Felten and Ralph Jozefowicz for the final examination in the University of Rochester first-year Neural Science course, illustrates one strategy to avoid hinging.

Each of the first three items focuses on a different aspect of the patient presentation, and students are likely to respond correctly to some and incorrectly to others, receiving “partial credit” for partial knowledge. The last item is probably slightly hinged on the preceding items, since it requires students to “put the whole picture together” in order to respond correctly, but this seems reasonable, given the importance of the latter.

It can be difficult for a single faculty member to prepare case clusters where the items draw on information from several basic science disciplines — this requires substantial breadth of knowledge. One strategy for coping with this problem is to adopt a “team approach” to preparation of test material analogous to the method generally used for preparation of problems for use in PBL instruction. For example, a clinician member of a team can prepare the patient description with which the cluster begins, along with questions related to pathophysiology. Faculty members from relevant basic science disciplines can contribute items that address various aspects of the patient situation from the perspective of their discipline.

Use of this kind of material is not, of course, restricted to curricula and courses taught using a PBL approach. It is completely appropriate any time it is desirable to stress clinical application of basic science information in teaching, learning and assessment. In our view, this includes most basic science courses — even those taught in the first year. As the neural science example on the next page illustrates quite well, it is straightforward and appropriate to test basic knowledge of anatomy and physiology in the context of patient care in a traditionally taught course.

An unresponsive 58-year-old woman is brought to the emergency department after collapsing at a local shopping mall. Her family reports that she felt well that morning but developed a headache that progressively worsened while she was shopping. She has had hypertension and atrial fibrillation and is taking an antihypertensive medication and an oral anticoagulant. Her blood pressure is 220/130 mm Hg and her respiratory pattern is one of apnea alternating with hyperpnea. She responds only to noxious stimuli with extensor posturing involving the right arm and leg. Fundoscopic examination reveals papilledema involving the left optic disc. Pupils are 3.0/7.0 (R/L) with no reaction to light on the left. There is a left gaze preference. There is diffuse hyperreflexia (R > L) and Babinski's sign is present bilaterally.

1. The dilated, unreactive left pupil is most consistent with injury to the left
 - A. optic nerve
 - B. optic tract
 - *C. oculomotor nerve
 - D. lateral geniculate nucleus
 - E. superior colliculus
2. The extensor posturing on the right is most consistent with injury to the left
 - A. telencephalon
 - B. diencephalon
 - *C. midbrain
 - D. pons
 - E. medulla
3. Her respiratory pattern is best described as
 - A. normal
 - *B. Cheyne-Stokes
 - C. central neurogenic hyperventilation
 - D. apneustic
 - E. ataxic
4. Which of the following herniation syndromes is most consistent with her clinical presentation?
 - A. Cingulate gyrus beneath the falx
 - *B. Temporal lobe uncus across the tentorium
 - C. Diencephalon through the tentorial notch
 - D. Brain stem through the tentorial notch
 - E. Cerebellar tonsils through the foramen magnum

Additional discussion of assessment in PBL courses and curricula can be found in:

Swanson DB, Case SM, and van der Vleuten CM. Strategies for student assessment. In: Boud, Feletti, eds. *The Challenge of Problem-Based Learning - Second Edition*. London: Kogan Page Ltd; 1997:269-282.

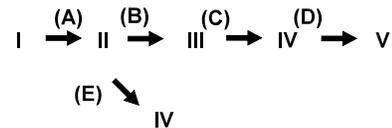
Sample Items for the Basic Sciences

- Several contiguous cells are labeled with a fluorescent dye that cannot cross cell membranes. One cell is experimentally bleached with light that destroys the dye, but soon recovers dye fluorescence. This recovery is best explained by the presence of which of the following structures between the bleached cell and its fluorescent neighbors?
 - A basal lamina
 - Desmosomes (maculae adherentes)
 - *C. Gap junctions
 - Glycosaminoglycans
 - Tight junctions (zonulae occludentes)

- A 30-year-old man has loss of pain and temperature sensation from the neck down on the right side of the body and on the left side of the face; partial paralysis of the soft palate, larynx, and pharynx on the left; and ataxia on the left. This syndrome is most likely to result from thrombosis of which of the following arteries?
 - Basilar
 - Right posterior inferior cerebellar
 - *C. Left posterior inferior cerebellar
 - Right superior cerebellar
 - Left superior cerebellar

- During an operation, the arterial PCO_2 and pH of an anesthetized patient are monitored. The patient is being ventilated by a mechanical respirator, and the initial values are normal ($\text{PCO}_2 = 40 \text{ mm Hg}$; $\text{pH} = 7.42$). If the ventilation is decreased, which of the following is most likely to occur?

Arterial PCO_2 pH	
A. Decrease	decrease
B. Decrease	increase
C. Decrease	no change
*D. Increase	decrease
E. Increase	increase
F. Increase	no change



- In the branched metabolic pathway, a different single enzyme catalyzes each of the individual steps. The enzyme that would be expected to be most severely inhibited by compound V is enzyme
 - A
 - *B. B
 - C
 - D
 - E

5. A patient with posthepatic cirrhosis develops rapid enlargement of the liver associated with deterioration of hepatic function. Serum concentration of which of the following is most likely to be abnormal?
- α_1 -Antitrypsin
 - Carcinoembryonic antigen
 - Chorionic gonadotropin
 - * α -Fetoprotein
 - Gastrin
6. The first-born infant of an Rh-negative 26-year-old woman who had two previous second trimester abortions has severe hemolysis and circulatory failure. This condition could have been prevented by treating the mother with
- anti-RhD IgG during the most recent pregnancy
 - *anti-RhD IgG on termination of each of the first two pregnancies
 - anti-RhD IgM during the most recent pregnancy
 - anti-RhD IgM on termination of the first pregnancy
7. Laboratory tests on an edematous 35-year-old man show a normal serum concentration of complement and an increased serum concentration of cholesterol. Urinalysis shows 4+ protein, 0-5 erythrocytes/hpf, and several hyaline casts. Examination of tissue obtained on renal biopsy is most likely to show
- acute poststreptococcal (proliferative) glomerulonephritis
 - membranoproliferative glomerulonephritis
 - *membranous glomerulonephritis
 - minimal change disease (lipoid nephrosis)
 - rapidly progressive glomerulonephritis
8. Genes on the bacterial chromosome have the following linkages in conjugal transfer: x and y, 25% of the time; y and z, 50% of the time. If the gene order is x-y-z, approximately what percentage of the time will x and z be transferred together?
- 1% of the time
 - 5% of the time
 - *13% of the time
 - 20% of the time
 - 40% of the time

9. At a banquet, the menu included fried chicken, home-fried potatoes, peas, chocolate eclairs, and coffee. Within 2 hours, most of the diners became violently ill, with nausea, vomiting, and abdominal pain. Analysis of the contaminated food is most likely to yield large numbers of which of the following organisms?
- A. *Escherichia coli*
 - B. *Proteus mirabilis*
 - C. *Salmonella typhimurium*
 - *D. *Staphylococcus aureus*
 - E. *Streptococcus faecalis*
10. Drug Y has a volume of distribution (Vd) of 75 L in both younger and older adult men. In younger adults, it has a clearance rate of 15 L/h, 50% of which is via the liver and 50% via the kidneys. For younger men, the maintenance regimen is 100 mg every 6 hours. Which of the following regimens will produce essentially the same steady-state concentration in an older man, whose creatinine clearance is reduced to half that of younger men, but whose hepatic function is unimpaired?
- A. 75 mg every 3 hours
 - *B. 75 mg every 6 hours
 - C. 75 mg every 9 hours
 - D. 100 mg every 3 hours
 - E. 100 mg every 6 hours
 - F. 100 mg every 12 hours
11. A patient seen in the emergency department does not know which “heart drug” he is taking. His heart rate is greater than 80/min, and the PR and QRS intervals on an ECG are prolonged. The patient reports ringing in his ears. Which of the following drugs has the patient most likely been taking?
- A. Digoxin
 - B. Lidocaine
 - C. Phenytoin
 - D. Propranolol
 - *E. Quinidine
12. An 8-year-old boy needs to be coaxed to go to school, and often, while there, he complains of severe headaches or stomach pain. Sometimes his mother has to take him home because of his symptoms. At night, he tries to sleep with his parents. When they insist he sleep in his own room, he says there are monsters in his closet. These findings are most consistent with which of the following diagnoses?
- A. Childhood schizophrenia
 - B. Normal concerns of latency-age children
 - *C. Separation anxiety disorder
 - D. Socialized conduct disorder
 - E. Symbiotic psychosis

Chapter 5

Item Content:

Testing Application of Clinical Science Knowledge



Methods for Assessment

Despite continued debate about the appropriateness of multiple-choice tests, all three Step exams of the USMLE continue to include multiple-choice questions (MCQs). In a quest for improved assessment instruments, the NBME has conducted continual research on other testing formats. For the past 25 years, a major focus of this research has been the Computer Based Examination (CBX) project, now termed Computer-based Case Simulations (CCS), which were introduced as a component of Step 3 in 1999. Since the mid-1970s, a second area of research has focused on standardized patients (SPs), which are scheduled for inclusion in the licensing examination sequence in 2004 or 2005. As with other forms of more “authentic assessment,” CCS and SP-based examinations appear to have significant advantages for assessment of aspects of clinical competence because they pose tasks for the examinee in a realistic way.

Other projects have focused on enhancing the multiple-choice format. As a result of test development research, MCQs today appear very different from those used in the past. For both content and psychometric reasons, true/false MCQ formats, such as K-types (multiple true/false) and C-types (A, B, Both, Neither), are no longer used on the licensure exams. While most of the questions on Step 2 have the traditional five options, both A-type questions and extended-matching questions can include as many as 26 options, pushing the examinee task to something closer to uncued free-response. Virtually every item on Step 2 provides a patient vignette that focuses on a task that is relevant to a new intern, such as determining the diagnosis or the next step in patient care. These items require interpretation and synthesis of the data that are provided; they also require application of knowledge to familiar or unfamiliar situations (depending on the experiences of the examinee). Like CCS and SP-based cases, modern multiple-choice items framed as brief patient vignettes present examinees with low-fidelity simulations of medical decision-making problems.

General Issues Regarding What to Test

There are several tensions that influence the construction of each Step exam which may be relevant to you as you consider what to include on your exams. As dictated by the purpose statements, the USMLE examination system is designed to be an examination for general licensure; Step 2, for example, is designed to assess the application of knowledge required for beginning the first post-medical school year, regardless of specialty. The focus on content that is necessary for practice rather than explicitly on content that has been taught (if it were possible to catalog such a body of information) means that questions might be included on the exam that assess knowledge not uniformly taught in medical school. Conversely, topics taught in some medical schools might be omitted from the examination. In reality, there is a close association between what is taught and what is included on the examination, but the examination is designed to be an independent assessment of what students need to know as they begin their internship, regardless of whether or not it is taught. The analogous issue for individual schools is the extent to which it is appropriate to include content on an exam that has not been explicitly “taught.” For example, you need to decide to what extent you want to hold students responsible for independent learning beyond what has been covered in lectures.

A second tension on Step 2 arises from the focus on knowledge that is required by the general undifferentiated physician. When critics have questioned the appropriateness of a general examination in an era of specialization, we have countered by noting that the medical license is a general license, not a specialty-specific license, which we believe mandates that the examination be general. The generalist initiative has reduced comment in this area and has made us appear somewhat prescient — not something the NBME is often accused of. The analogous issue for individual schools is the extent to which clerkships should focus exclusively on “their” discipline, without an attempt to integrate across disciplines. In addition, faculty must decide on the extent to which clerkships should serve as a “pre-internship” experience for students, as opposed to providing an overview of topics that are relevant to all practicing physicians regardless of specialty.

The focus on questions that are appropriate for all graduating students, rather than those who are entering specific specialties has had a considerable influence on the content of Step 2 items. For example, the surgery committee is more likely to write items on patient management situations where immediate intervention is essential, or where it is important to triage the patient to a surgical floor rather than discharge him from the emergency department than to write questions related to surgical techniques that would not be considered essential knowledge for non-surgeons. The obstetrics and gynecology committee is less likely to write questions on the management of unusual high-risk pregnancies (even if most students have been exposed to these cases in the clerkship) than on the diagnosis of ectopic pregnancy — something every new intern, regardless of specialty, should recognize. Psychiatry items are unlikely to cover inpatient psychiatry but are more likely to assess ability to differentiate psychiatric from nonpsychiatric illnesses in patients who present with bizarre behavior. All questions

on Step 2 must be accepted by the entire discipline-specific committee; in addition, they must be accepted by an interdisciplinary committee. This discipline-specific review followed by interdisciplinary review helps to achieve content balance on the exam. In your own schools, it seems important to have test questions reviewed by other members of your own department. It is likely to be informative to have test questions reviewed by faculty members outside your department, though the logistics of implementing this procedure may be daunting.

Testing Recall of Isolated Facts or Application of Knowledge

In addition to thinking about the topics that are important to include on a test, you should think about how to structure those questions to test more than recall of isolated facts. Traditionally, test questions have been classified as requiring recall, interpretation, or problem solving (memory, comprehension, and reasoning) depending on the cognitive processes required to answer the question. Typical definitions refer to “Recall Questions” as those which assess examinee knowledge of definitions or isolated facts. “Interpretation Questions” require examinees to review some information, often in tabular or graphic form, and reach some conclusion (eg, a diagnosis). “Problem-Solving Questions” present a situation and require examinees to take some action (eg, the next step in patient management). The difficulty with these classifications is that the cognitive processes required to answer a question are as dependent on the background of the examinee as they are on the question content. Experts in a content area may simply recall an answer with little or no conscious thought, whereas others may need to reason out the answer from basic principles. The cognitive processes involved in responding to a question are examinee-specific, making the taxonomic approach difficult to use.

An alternate approach divides items into two categories: application of knowledge or recall of an isolated fact. If a question requires an examinee to reach a conclusion, make a prediction, or select a course of action, it is classified as an *application of knowledge* question. If a question assesses only rote memory of an isolated fact (without requiring its application), it is classified as a recall question.

Questions asking for recall of isolated facts often begin by citing a disease and then asking what patient findings are expected. These questions are structured similarly to most textbooks; the examinee could look up the disease and find the answer in a single paragraph. The flaw with these items is that they seem clinically backward. Patients rarely tell their physician what disease they have and then ask the physician what their signs and symptoms are.

Which of the following findings is most likely to be seen in postsurgical patients with pulmonary embolism?

Another type of question posing an inappropriate task is referred to as a “waiting room item.” In this item, the examinee is asked to select one of five patients for whom fetal karyotyping is most appropriate, almost as if the examinee is charged with performing fetal karyotyping on someone and he has only to look into the waiting room and select the patient who is most appropriate.

In contrast, the following item describes a patient and asks which study is most appropriate, a more reasonable task.

Questions assessing application of knowledge are structured in a clinically more realistic manner, for example, by giving the findings and asking the examinee to indicate the underlying disease. Typically, examinees would need to be able to synthesize information from several pages of a textbook to answer these questions.

Which of the following is an indication for fetal karyotyping in a 28-year-old woman?

- A. Paternal age 55
- B. Fetal cystic hygroma on ultrasound exam
- C. Previous child with spina bifida
- D. Previous miscarriage of a triploid fetus
- E. Trisomy 21 in the woman's brother

A healthy 28-year-old teacher is at 11 weeks' gestation. Family history is unremarkable except that both of her brothers have severe mental retardation, her mother died at 55 years of age of breast cancer, and her father is estranged. No family health records are available. Which of the following studies is appropriate?

- A. Blood test for fragile X carrier status
- B. Blood test for phenylketonuria carrier status
- C. Chorionic villus sampling for Duchenne's muscular dystrophy
- D. Chorionic villus sampling for chromosome analysis
- E. Amniocentesis for "-fetoprotein

A 62-year-old man develops acute shortness of breath and pleuritic chest pain 4 hours after undergoing cholecystectomy. Which of the following is the most likely diagnosis?

Use of questions with clinical vignettes as item stems has several benefits. First, the “face validity” of the examination is greatly enhanced by using questions that require examinees to “solve” clinical problems. Second, questions are more likely to focus on important information, rather than trivia. Third, these questions help to identify those examinees who have memorized a substantial body of factual information, but are unable to use that information effectively in clinical situations. Questions with clinical vignettes are generally more appropriate for higher level examinations.

Writing application of knowledge questions is relatively straightforward in medicine. When you describe a patient and ask a question related to that patient, you are assessing application of knowledge. The one instance in which use of a clinical vignette involves simply recall of an isolated fact is if the vignette describes a patient identical to one the student has read about before (eg, you use a patient vignette from a textbook or one discussed in class).

Guidelines for Clinical Science Item Content

- Test application of knowledge using clinical vignettes to pose medical decisions in patient care situations.
- Focus items on common or potentially catastrophic problems; avoid “zebras” and esoterica.
- Pose clinical decision-making tasks that would be expected of a successful examinee.
- Avoid clinical situations that would be handled by a (sub)specialist.

Questions should focus on specific tasks that the successful examinee must be able to undertake at the next stage of training (eg, determine the most likely diagnosis; indicate what additional laboratory studies should be ordered; formulate the next step in management; predict the most likely additional finding). For each topic, the areas in which mistakes are commonly made should be the focus of a question.

Writing One-Best-Answer Items

Constructing the Stem

The vast majority of questions should be written with a clinical vignette. The stem should begin with the presenting problem of a patient, followed by the history (including duration of signs and symptoms), physical findings, results of diagnostic studies, initial treatment, subsequent findings, etc. Vignettes may include only a subset of this information, but the information should be provided in this specified order. The stem should consist of a single, clearly formulated problem. The lead-in of the stem must pose a clear question so that the examinee can pose an answer without looking at the options. Satisfying the “cover-the-options” rule is an essential component of a good question.

Good stem: This stem 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. His temperature is 37.2 C (99 F). Breath sounds are distant with a few rhonchi and wheezes. His leukocyte count is 9000/mm³ with a normal differential. Gram's stain of sputum shows numerous neutrophils and gram-negative diplococci. X-ray films of the chest show hyperinflation. Which of the following is the most likely diagnosis?

Stem testing isolated facts: The following stem contains insufficient information; in order to answer the question, the examinee must use the options as a frame of reference.

Which of the following is true about pseudogout?

Patient vignettes should include some or all of the following components in the order indicated:

- Age, Gender** (eg, A 45-year-old man)
- Site of Care** (eg, comes to the emergency department)
- Presenting Complaint** (eg, because of a headache)
- Duration** (eg, that has continued for 2 days).
- Patient History (with Family History ?)**
- Physical Findings**
- +/- Results of Diagnostic Studies**
- +/- Initial Treatment, Subsequent Findings, etc.**

Make sure that your stem:

- 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 (ie, avoid using *except* or *not* in the lead-in)

Fine Points on Item Stems

Use of Real Patients. We believe it is generally better not to base multiple-choice questions on “real patients,” particularly for tests aimed at students. As a general rule, real patients are too complicated, and the elements that are complicated are not necessarily those that are important for assessment. As noted earlier, we do include window dressing (ie, incidental findings), but do not include “red herrings” (ie, information that is intended to lead examinees away from the correct answer). Unfortunately, real patients often have “red herrings” among their findings.

Use of Reference Materials. We believe that it is appropriate to provide information in a test question if, in real life, someone would be likely to refer to a reference source to obtain the information. For example, in many instances, we believe it is appropriate to provide a table of normal laboratory values or a chart showing a recommended schedule of screening tests or immunizations. Of course, you might not just ask questions that require examinees to simply look up information in the chart provided, but you might, for example, ask about immunization of a 6-year-old child who had never been immunized.

Use of Patient’s or Physician’s Own Words. We generally do not believe it is useful to include the patient’s own words, particularly if the examinee task is to interpret nuances of language that might be affected by tone. On the other hand, it may be useful to ask the examinee to select the most appropriate physician response to a patient by asking the examinee to choose among options phrased as open-ended, closed, or leading questions.

Patients Who Lie. We believe all multiple-choice patients should tell the truth, or the physician’s interpretation of the patient’s story should be provided. Physicians use multiple cues to determine how truthful a patient is and many of these cues cannot be translated into written form. Thus, our items would describe a patient’s alcohol consumption as “The patient drinks 16 oz of beer with dinner each night” or “The patient’s description of his alcohol consumption is contradictory.” We would not write something ambiguous, such as “The patient ‘claims’ to drink only one bottle of beer each night.”

Verbosity, Window Dressing, and Red Herrings: Do They Make a Better Test Item?

Most educators stress the importance of writing item stems that are as short as possible, avoiding verbosity (ie, extra words), “window dressing” (ie, extraneous material), and “red herrings” (ie, information designed to mislead the examinee). Somewhat in opposition to this advice, we have emphasized use of clinical vignettes in item writing efforts. For USMLE Step 2, these vignettes consist of paragraph-length descriptions of clinical situations, generally followed by a question related to the diagnosis or next step in patient care. Such items stress application of knowledge by asking examinees to make clinical decisions, rather than to simply recall isolated facts. They are designed to reflect “real life tasks” by challenging examinees to first identify the findings that are important, then integrate those findings into a diagnosis or clinical action. Such items often require multiple steps in the thinking process. We have found that vignette items tend to have fewer technical item flaws than typical non-vignette items, presumably because vignettes follow a standard structure and pose questions that are clinically natural.

Despite these advantages, some have challenged the use of vignettes, believing that a vignette only makes an item more cumbersome by wrapping window dressing around the real question. Some advocate avoiding vignettes altogether; others advocate using short vignettes and including only relevant positive findings expressed concisely; the rest advocate use of long vignettes that include more complete information that the examinee must review and synthesize.

Several studies¹ were conducted to compare the psychometric characteristics of items developed in three formats: non-vignette, short vignette, and long vignette. The progression was designed to require increasing levels of interpretation, analysis, and synthesis of findings (see sample item in three formats below). As expected, items became more difficult as patient findings were presented in a less interpreted form; however, the differences in discrimination were not statistically significant. Regardless of the mixed psychometric results, we believe vignette items are generally more appropriate because they test application of knowledge to patient situations and pose appropriate clinical challenges; such items might be viewed as “low fidelity” clinical simulations that improve the content validity of the examinations.

An item written in a non-vignette format typically is written from a “top-down” perspective (ie, given a disease, what are the associated findings). To an expert, items written in this manner may appear identical to items written with a patient vignette.

¹From: Case SM, Swanson DB, Becker DF. Verbosity, window dressing, and red herrings: do they make a better test item? *Academic Medicine*. 1996;71:528-530.

The following trio of items indicate that, to the low-performing student, the items are not identical.

Each item was administered to senior medical students as part of the licensing examination. The grid under each item shows the percentage of “Hi” and “Lo” students (ie, those who performed in the top and bottom 20% of the examination overall) who selected each option. Almost all of the Hi group (99%) and the Lo group (90%) selected the correct option in the non-vignette format. The short- and long-vignette formats were not markedly more difficult for the Hi group, but were for the Lo group; the correct answer was selected by 82% in the short-vignette format and 66% in the long-vignette format.

Non-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
Hi	1	0	99	0	0
Lo	8	1	90	1	0

Short Vignette

A 2-year-old boy has a 1-week history of edema. His blood pressure is 100/60 mm Hg, and there is a 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. The most likely diagnosis is (same option list followed).

	A	B	C	D	E
Hi	0	0	98	2	0
Lo	5	2	82	8	1

Long Vignette

A 2-year-old black 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. (Same lead-in and option list followed).

	A	B	C	D	E
Hi	0	1	98	1	0
Lo	10	9	66	10	5

Although the item listed above is labeled “long vignette,” it is quite short in comparison to most items on USMLE Step 2. We believe that new graduates from medical school need to demonstrate proficiency in sorting through patient information, synthesizing the important findings, and reaching a conclusion. As a result, our items tend to have a mix of important and unimportant findings. On the other hand, we sometimes synthesize findings with a statement such as “the family history is noncontributory.”

Writing Items Related to Physician Tasks

Each topic requires a slightly different approach to item writing. We provide some sample lead-ins and other tips to guide your item-writing efforts for each physician task.

Health and Health Maintenance

Items in this topic area assess ability to evaluate risk factors, understand epidemiologic data, and apply preventive measures. Health and health maintenance items commonly fall into one of the following categories: 1) screening tests; 2) constructive interference; 3) immunizations/travel medicine; or 4) emergency intervention.

In general, begin the items with a clinical vignette that describes a patient. In addition to history and physical examination findings, these vignettes may include information about immunization 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 applications of this knowledge. In asking about immunizations or screening tests, consider providing a chart of customary practices to avoid memorization of conflicting recommendations.

The following lead-ins are examples of those used in this category:

Which of the following immunizations should be administered 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 the 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 should be recommended to prevent disability from this injury/condition?

Early treatment with which of the following is most likely to have prevented this patient's condition?

Supplementation with which of the following is most likely to have prevented this condition?

A 15-year-old boy has had two episodes of severe anaphylactic shock following bee stings. Which of the following is the most appropriate intervention?

- A. Administration of corticosteroids during the summer
- B. Long-term prophylactic antihistamine therapy
- C. Protective clothing
- *D. Desensitization with bee venom extract
- E. Restrict him to the house during the summer months

A 33-year-old woman, gravida 1, para 1, spontaneously delivers a 2460-g (5 lb 7oz) 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. 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

An asymptomatic 33-year-old man has a blood pressure of 166/112 mm Hg. Serum electrolyte levels are within normal limits. Effective antihypertensive treatment is most likely to reduce the likelihood of which of the following?

- A. Aortic aneurysm
- B. Congestive heart failure
- C. Myocardial infarction
- D. Renal failure
- *E. Stroke

Mechanisms of Disease

These items should test examinees' 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. Mechanisms questions should be framed in a clinical context.

In general, begin your mechanisms items with a clinical vignette of a patient and his/her symptoms, signs, history, lab study findings, etc. Then ask a question such as one of these:

Which of the following is the most likely explanation for these findings?

Which of the following is the most likely location of the patient's lesion?

Which of the following is the most likely pathogen?

Which of the following findings is most likely to be increased/decreased?

A biopsy is most likely to show which of the following?

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 (BUN) level is 3.2 mg/dL. Which of the following is the most likely cause?

- A. Coarctation of the aorta
- B. Decreased production of endothelial-derived relaxant factor
- C. Increased production of aldosterone
- D. Increased production of catecholamine
- *E. Intravascular volume of expansion

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 are HIV negative. Examination of the patient's sexual partner shows an anal fissure; urethral culture does not grow N. gonorrhoeae. Which of the following is the most likely cause of the recurrence of urethral infection?

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

Diagnosis

For this category, write items that require examinees to interpret history, physical findings, and results of laboratory, imaging, and other studies in order to determine the most likely diagnosis (differential diagnosis) or the next step in diagnosis (diagnostic testing). As you write questions in this area, think about whether you want to assess ability to integrate knowledge across clerkships.

The classic diagnosis item begins with a patient description (including age, gender, symptoms and signs and their duration, history, physical findings on exam, findings on diagnostic and lab studies) and ends with a question:

Which of the following is the most likely diagnosis?

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

Which of the following is most likely to confirm the diagnosis?

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. His temperature is 37.2 C (99 F). Breath sounds are distant with a few rhonchi and wheezes. His leukocyte count is 9000/mm³ with a normal differential. Gram's stain of sputum shows numerous neutrophils and gram-negative diplococci. X-ray films of the chest show hyperinflation. Which of the following is the most likely diagnosis?

- A. Asthma
- B. Bronchiectasis
- *C. Bronchitis
- D. Pulmonary embolism
- E. Streptococcal pneumonia

A 28-year-old woman has palpitations that occur approximately once a week, last 1-5 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 coffee daily. She rarely drinks alcohol and does not smoke. 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 left upper sternal border. An ECG is normal except for evidence of sinus tachycardia. Which of the following is the most appropriate next step in diagnosis?

- A. Ambulatory ECG monitoring
- B. Measurement of serum thyroid-stimulating hormone level
- C. Measurement of urine catecholamine level
- D. MUGA scan
- E. Echocardiography

Management

These items assess principles of chronic and acute care in inpatient and outpatient settings. When writing management items, it is especially important to focus on aspects of care relevant to the level of practice of the examinee (supervised, limited supervision, independent practice, subspecialist). For Step 2, we focus on aspects of care that would be appropriate for all new interns regardless of specialty, and middle-of-night urgent care when help may be unavailable. Questions cover pharmacotherapy decisions, more general treatment decisions where options include a mix of pharmacotherapy and other treatments, and management decisions where options include a mix of treatments and other things such as diagnostic testing.

Questions to ask include:

Which of the following is the most appropriate initial or next step in patient care?

Which of the following is the most effective management?

Which of the following is the most appropriate pharmacotherapy?

Which of the following is the first priority in caring for this patient? (eg, in emergency department)

A hospitalized 55-year-old woman with decompensated cirrhosis of the liver is being treated with spironolactone, potassium chloride elixir, and furosemide. She is now barely responsive and hypotensive without respiratory distress. She has signs consistent with chronic hepatic disease, ascites, and minor peripheral edema. ECG shows a regular, slow (55/min) rhythm, no P waves, and a wide, slurred QRS complex running into a wide, slurred ST and T wave. Which of the following should be administered intravenously?

- A. Calcium
- B. Lidocaine
- C. Magnesium
- D. 0.9% Saline
- *E. Potassium

A previously healthy 15-year-old boy has cramping periumbilical pain; after several hours, the pain shifts to the right lower quadrant and becomes constant. He vomits several times and is brought to the emergency department. The abdomen is tender on deep palpation of the right lower quadrant. Findings on chest and abdominal x-ray films are normal. Leukocyte count is 15,000/mm³. Urinalysis shows 3 leukocytes/hpf. Which of the following is the most appropriate initial management?

- A. Supportive treatment at home; return at once if the pain increases
- B. Barium enema
- C. CT scan of the abdomen
- D. Intravenous pyelography and cystography
- *E. Surgical exploration of the abdomen

Writing Items on Difficult Topics

One of the common beliefs is that many topics do not lend themselves to a multiple-choice format. We have had reasonable success in generating items in many of these areas and recommend the following strategy.

1. After you identify a topic that you're having difficulty with, look through all sources of test material and select any questions on the topic that you think are acceptable.
2. Identify the key features of these items and try to develop a template that would enable faculty to write similar items.
3. For topics where no sample items are available, think about what you want to assess. Go beyond the list of topics by outlining tasks related to the topic that are essential for medical students to know (ie, each element of the list should include a verb).

To illustrate this process, the following paragraphs outline the process we used to write Step 2 questions on Medical Ethics and Jurisprudence. The content outline includes the following topics: 1) Consent and informed consent to treatment (eg, full disclosure, alternate therapies, risks and benefits); 2) Physician-patient relationship (eg, truth-telling, confidentiality, privacy, autonomy, public reporting); 3) Death and dying (eg, diagnosing death, life-support, autopsy, organ donation, euthanasia, suicide); 4) Birth-related issues (eg, prenatal diagnosis, abortion, maternal-fetal conflict); and 5) Research issues (eg, consent, placebos, conflict of interest, vulnerable populations).

Authors typically wrote questions such as: Which of the following is the definition of informed consent? Or they would describe a scenario and ask which ethical principle was illustrated by the scenario. This is a sample legal question from the past.

These are irreverently referred to as “who cares” questions.

The legal basis for the eased restrictions on abortions in the US can be traced most closely to

- A. *federal legislation*
- B. *a federal court ruling*
- C. *state legislations*
- D. *state court rulings*
- E. *AMA rulings*

We reviewed the item pool and decided that we were less interested in whether or not students knew the definitions; we wanted to assess whether or not examinees could apply ethical principles in their decisions related to patient care. We then convened a group of item writers who looked through model questions and generated new items for the exam. All questions involved a patient vignette and asked the student to indicate what the physician should do or asked the student to evaluate the appropriateness of the physician's actions indicated in the vignette; no questions focused on the definition of terms. The following are some sample items we wrote.

A nurse is hospitalized for an appendectomy at the medical center where she is employed. One week after discharge, the assistant hospital administrator asks the surgeon what the final diagnosis was. Which of the following is the most appropriate response on the part of the surgeon?

- A. *Answer, because it will expedite handling of insurance issues at the medical center*
- B. *Answer, because as an employee of the medical center the administrator has access to information about patients*
- C. *Answer, because of the possibility of spreading misinformation about the patient*
- D. *Decline to answer, because the administrator is not a medical doctor*
- *E. *Decline to answer, because the information is confidential*

An 8-year-old boy with acute lymphoblastic leukemia has experienced three relapses in the past 2 years. The only available treatment is experimental chemotherapy. Without treatment, the child is unlikely to survive for more than 6 weeks; with treatment, his prognosis is unknown. The parents do not want further treatment for their son and wish to take him home; the child also says he wants to go home. Which of the following is the most appropriate course of action?

- A. *Discharge the child against medical advice*
- *B. *Discharge the child routinely*
- C. *Petition the court for an order for treatment*
- D. *Report the parents to social services for medical neglect*

Section III

Extended-Matching Items



Section III provides information regarding Extended-Matching items. Chapter 6 presents the Extended-Matching item format where examinees are instructed to select the one-best answer. Chapter 7 presents the format where examinees are instructed to select some particular number of options, generally more than one.

Chapter 6

Extended-Matching (R-Type) Items



Extended Matching items are multiple-choice items organized into sets that use one list of options for all items in the set. A well-constructed Extended-Matching set includes four components:

1. a theme;
2. an option list;
3. a lead-in statement; and
4. at least two item stems, as illustrated below.

Theme: **Fatigue**

Options:

A. Acute leukemia	H. Hereditary spherocytosis
B. Anemia of chronic disease	I. Hypothyroidism
C. Congestive heart failure	J. Iron deficiency
D. Depression	K. Lyme disease
E. Epstein-Barr virus infection	L. Microangiopathic hemolytic anemia
F. Folate deficiency	M. Miliary tuberculosis
G. Glucose 6-phosphate dehydrogenase deficiency	N. Vitamin B ₁₂ (cyanocobalamin) deficiency

Lead-in: For each patient with fatigue, select the most likely diagnosis.

Stems:

1. A 19-year-old woman has had fatigue, fever, and sore throat for the past week. She has a temperature of 38.3 C (101 F), cervical lymphadenopathy, and splenomegaly. Initial laboratory studies show a leukocyte count of 5000/mm³ (80% lymphocytes, with many lymphocytes exhibiting atypical features). Serum aspartate aminotransferase (AST, GOT) activity is 200 U/L. Serum bilirubin concentration and serum alkaline phosphatase activity are within normal limits.

Ans: E

2. A 15-year-old girl has a two-week history of fatigue and back pain. She has widespread bruising, pallor, and tenderness over the vertebrae and both femurs. Complete blood count shows hemoglobin concentration of 7.0 g/dL, leukocyte count of 2000/mm³, and platelet count of 15,000/mm³.

Ans: A

Extended-matching items are written differently than traditional one-best-answer items. Most often, the theme, lead-in, and options are written first; the item stems are written last. For example, if you want to write some questions related to the diagnosis of fatigue, you would begin by listing the diagnoses that might cause fatigue. You would then write a vignette for each (or many) of the options in the list. The example above includes vignettes for Epstein-Barr virus infection and for acute leukemia. Additional items might be written for some of the remaining diagnoses; for common, treatable diagnoses, more than one item might be prepared. The sample vignettes are moderate in length; shorter, more-focused vignettes could also be used. Alternatively, examinees could be challenged to identify key diagnostic information, intermingled with incidental findings, by using longer vignettes.

Avoiding Flaws When You Write Extended-Matching Items for Your Own Examination

The four components (theme, options, lead-in, and stems) are all essential for the construction of a good quality extended-matching set. Sets without lead-ins (or with nonspecific lead-ins, such as “Match each item with the best option”) should NOT be used, because they generally pose inconsistent or ambiguous tasks for examinees. The following set is flawed. The options are heterogeneous; there is no lead-in; the stems cannot be answered without reading the options. Rules for extended-matching items are completely analogous to those for one-best answer items.

Sample Extended-Matching Set - Flawed

- | | |
|--|---------------------------------|
| A. is motion sickness | I. are completely controlled |
| B. have no effects on people | J. cause plant and eye damage |
| C. indirectly increase CO ₂ | K. are negligible |
| D. cause death | L. increase risk of skin cancer |
| E. increased odor sensitivity | M. cannot be controlled |
| F. is a reduction in visibility | N. excess acute respiratory |
| G. esthetics, economics, health | illness among children |
| H. products of fossil fuel combustion | O. contrary to public opinion |
-
1. Factors that people consider when evaluating air quality
 2. A principal effect of particulate matter in air
 3. The products of photochemical smog

After reading the stem in Item #1, examinees have only the vaguest idea what the question is about. In an attempt to determine the “best” answer, the examinees have to decide whether “is motion sickness” is more or less true than “have no effects on people.” The task is not do-able. Under these circumstances, unless an option is absolutely 100% true or false, it cannot be rank-ordered with the other options. The stem of item #1 by itself is not clear; the item cannot be answered without looking at the options.

As with one-best-answer items, the stems should be long; the options should be short. There MUST be a lead-in that establishes the relationship between the items and the options. There should be NO verbs in the options. The “cover-the-options” rule is as relevant for extended-matching items as it is for one-best-answer items.

Sample Lead-ins and Topics for Option Lists

Patient vignettes provide an excellent structure for stems, not only in the clinical sciences, but also to assess knowledge in the basic sciences. Lead-ins generally begin with a phrase such as “For each of the following patients.” Often sets are organized around chief complaints or some other factor that allows a more specific introductory phrase such as “For each of the following patients with fatigue,” or “For each of the following patients with an enzyme deficiency.” The second part of the lead-in describes the task and the option set: “select the most likely diagnosis”; “select the protein that is most likely to be defective.”

The following are some additional sample lead-ins and some suggested topics for option lists.

- For each of the following patients, select the [eg, nerve] that is most likely to be [abnormal/defective/deficient/non-functioning].
Options sets could include list of nerves; list of muscles; list of enzymes; list of hormones; list of proteins; list of types of cells; list of neurotransmitters; list of pathologic processes.
- For each of the following patients, select the [finding] that would be expected.
Options sets could include list of laboratory results; list of additional physical signs; autopsy results; results of microscopic examination of fluids, muscle or joint tissue; DNA analysis results; hormone levels.
- For each of the following patients, select the most likely [cause].
Options sets could include list of underlying mechanisms of the disease; medications that might cause side effects; list of drugs or drug classes; toxic agents; hemodynamic mechanisms.
- For each of the following patients, select the [eg, drug] that should be administered.
Options sets could include list of drugs, vitamins, amino acids, enzymes, hormones.
- For each of the following patients with [chief complaint], select the most likely diagnosis.
Options sets could include list of diagnoses, most often organized around a chief complaint such as diseases that cause chest pain or diseases that cause fever.
- For each of the following patients, select the most appropriate next step in patient care.
Options sets could include list of pharmacologic therapies, list of laboratory studies, disposition alternatives, or the options could include a mixed set of treatments and additional studies to assess whether the student knows when sufficient data have been gathered.

More on Options for R-Sets

Generally, anything that can be listed can form the basis for options in an R-set. Below are some topics that have been used as the basis for option lists.

Arteries	Connective Tissue Types
Nerves	Anatomic Structures
Muscles	Endocrine Structures
Amino Acids	Neurotransmitters
Peptides	Metabolic Defects
Hormones	Immune Disorders
Enzymes	Motor System Components
Cell Components	Cardiac Structures
Cell Types	Organelles
Blood Components	Congenital Anomalies
Molecules	Segments of the Spinal Cord
Karyotypes	Central Nervous System Components
Proteins	Secretory Products
Lipids	Extracellular Matrix Components
Pathogens/Bacteria/Fungi	Management Alternatives
Viruses	Drugs/Drug Classes
Cytokines	Pathologic Processes
Toxins	Pathophysiologic States
Vitamins/Minerals	Electrolyte Abnormalities
Diagnoses	Diagnostic Tests

The list of options should be single words or very short phrases. They must be homogeneous (all diagnoses, all management options, all anatomical sites, all vitamins, etc). They can be labeled areas in a graph or in pictorial material. Options, especially those involving laboratory values, are often expressed in tabular form (see physiology example). Include all relevant options that are appropriate for the examinees; subtle distinctions and uncommon diagnoses may be inappropriate. For some topics, as few as three options might be appropriate; for others, a list of 26 (one for each letter in the alphabet) might be required.

Writing the Item Stems

Patient vignettes provide an excellent structure for stems both in the basic and the clinical sciences. In the clinical sciences, the vignette commonly provides the patient's age, gender, chief complaint, and site of care, followed by personal history, family history (if relevant), then physical examination information, then laboratory data (if provided). Depending upon the purpose of the set, vignettes can be brief, prototypic presentations or fuller descriptions that challenge examinees to identify key information. Generally, these items would include at least the patient's age, gender, chief complaint, and related history. Items to assess knowledge of the basic sciences, particularly for courses taught in the first year of medical school, might include less detailed prototypical descriptions.

Each patient description should be similar in structure to the others in the set. For example, if race, ethnicity, or occupation is included in one item, it should be included in all items; if laboratory data are included for one item, include them in all items. It is advisable not to mix adults and pediatric cases in the same set — too often the age alone provides too much cueing and eliminates large numbers of options from consideration.

One advantage of the use of patient vignettes is that it helps to assure that the content assesses application of knowledge. These items should not resemble crossword puzzles, where both the options and the stems are single words or short phrases. Avoid reconstructing those items you were faced with in junior high school where you had to draw a line from something in column A to the matching option in column B.

It is particularly important that the items be straightforward. There is no reason to make them tricky; the extended option list makes them difficult enough to allow you to distinguish the knowledgeable student from the other students without resorting to trickery. As with well-constructed A-type questions, the “cover-the-options” rule is paramount. Knowledgeable students should be able to generate an answer to the question and then find that answer in the alphabetical list of options.

An item should be prepared for most of the options; for common or important options, more than one item can be written. In constructing an examination assessing general competence, to avoid overemphasizing a topic, all of the options, but only two or three of the items, would be used; the remaining items are retained for subsequent exams. On the other hand, if you want to assess knowledge in greater depth on a smaller number of topics, 10 to 20 items can be included for each set, with a sub-score calculated for each topic.

In reviewing the items, check to make sure that there is only a single “best” answer for each question. Also make sure that there are at least four reasonable distractors for each item. As a final check, it is recommended that you ask a colleague to review the items (without the correct answer indicated). If the colleague has difficulty determining the correct answer, modify the option list or the item to eliminate the ambiguity.

Sample Good and Bad Item Stems Using the Same Option List

The following is a good microbiology set. The options are a homogeneous list of pathogens; including both viruses and bacteria makes sense. There is a lead-in that presents a clear task for the examinee. There are two item stems that require students to apply their basic science knowledge of microbiology to arrive at the most likely cause of each patient's illness.

- | | |
|---------------------------------------|--------------------------------------|
| A. Adenovirus | L. <i>Haemophilus influenzae</i> |
| B. <i>Aspergillus fumigatus</i> | M. <i>Histoplasma capsulatum</i> |
| C. <i>Bacillus anthracis</i> | N. <i>Mycobacterium tuberculosis</i> |
| D. <i>Candida albicans</i> | O. <i>Mycoplasma pneumoniae</i> |
| E. <i>Chlamydia psittaci</i> | P. <i>Neisseria gonorrhoeae</i> |
| F. <i>Coccidioides immitis</i> | Q. <i>Neisseria meningitidis</i> |
| G. Coronavirus | R. <i>Pneumocystis carinii</i> |
| H. <i>Corynebacterium diphtheriae</i> | S. Rhinovirus |
| I. <i>Coxiella burnetii</i> | T. <i>Streptococcus pneumoniae</i> |
| J. Coxsackievirus | U. <i>Streptococcus pyogenes</i> |
| K. Epstein-Barr virus | (group A) |

For each patient with fever, select the pathogen most likely to have caused his/her illness.

1. A 7-year-old girl has a high fever and a sore throat. There is pharyngeal redness, a swollen right tonsil with creamy exudate, and painful right submandibular lymphadenopathy. Throat culture on blood agar yields numerous small β -hemolytic colonies that are inhibited by bacitracin.

Ans: U

2. For the past week, an 18-year-old man has had fever, sore throat, and malaise with bilaterally enlarged tonsils, tonsillar exudate, diffuse cervical lymphadenopathy, and splenomegaly. There is lymphocytosis with atypical lymphocytes. The patient tests positive for heterophil antibodies.

Ans: K

The following stem, developed for the same set, assesses recall of isolated facts rather than application of knowledge. It looks more like a crossword puzzle question than a question for a medical school examination.

3. An encapsulated gram-positive organism that usually grows in pairs or short chains.

Ans: T

The following set includes two item stems. The first requires that the examinee synthesize information in order to determine a diagnosis; the second requires only recall of an isolated fact.

- | | |
|---------------------------|--------------|
| A. Vitamin A | I. Biotin |
| B. Vitamin B ₁ | J. Copper |
| C. Vitamin B ₂ | K. Folate |
| D. Vitamin B ₆ | L. Iodine |
| E. Vitamin C | M. Iron |
| F. Vitamin D | N. Magnesium |
| G. Vitamin E | O. Niacin |
| H. Vitamin K | P. Zinc |

For each patient with clinical features caused by metabolic abnormalities, select the vitamin or mineral that is most likely to be involved.

1. A 70-year-old widower has ecchymoses, perifollicular petechiae, and swelling of the gingiva. His diet consists mostly of cola and hot dogs.

Ans: E

2. Involved in clotting factor synthesis.

Ans: H

The following set includes two item stems. The first item stem requires that the examinee synthesize information to determine the diagnosis of trisomy 21; the second item provides this information. When you develop item stems, you need to decide the level of synthesis you will provide.

- | | |
|---|------------------------------|
| A. Atrial septal defect | E. Patent ductus arteriosus |
| B. Coarctation of the aorta | F. Pulmonic valve stenosis |
| C. Complete transposition of the great arteries | G. Tetralogy of Fallot |
| D. Endocardial cushion defect | H. Ventricular septal defect |

For each patient, select the most likely congenital heart defect.

1. A 3090-g infant is born at term. Physical examination shows the child to be hypotonic with a weak suck. There is no cyanosis. Prominent epicanthal folds, a large tongue, and small incurved 5th digits on the hands are present. There is a loud holosystolic murmur over the entire precordium and a palpable thrill at the upper left sternal border in the 2nd through 4th intercostal spaces.

Ans: D

2. An infant with trisomy 21 has clefts of the right and left atrioventricular valves; there is no cyanosis.

Ans: D

This set poses a task that is clinically backward. The examinee is given a vaccine and asked to select the best patient for this vaccine. A more appropriate task would be for the examinee to be given a patient and asked for the most appropriate next step in patient care (ie, describing a patient in each item stem; using “For each patient, select the most appropriate next step in patient care”; and using things such as vaccines as options). A second problem with the set is that insufficient information is provided about each patient. For example, the examinee would want the immunization history for a patient before deciding which immunizations to provide.

	Birth Year	Gender	Occupation	Pregnant	Childhood Disease	Medical History	Allergy
A.	1980	Male	Student	-	None	None	Egg products
B.	1975	Female	Painter	Yes	None	None	Penicillin
C.	1970	Female	Teacher	Yes	None	None	None
D.	1965	Female	Lawyer	No	None	None	Gelatin products
E.	1960	Male	Painter	-	Measles	None	Tetanus toxoid
F.	1955	Female	Clerk	No	Mumps	Diabetes	None
G.	1950	Female	Nurse	No	Varicella	None	None
H.	1945	Male	Executive	-	Measles	Hypertension	None
I.	1940	Male	Driver	-	Rubella	Splenectomy	None
J.	1935	Female	Homemaker	No	Unknown	Gastritis	Sulfonamides

For each vaccination, select the patient profile that represents its most appropriate use.

1. Measles vaccine
2. Meningococcal vaccine

Overview of the Steps for Writing Extended-Matching Items

1. **Identify the theme for the set.** The theme can be a chief complaint (eg, chest pain, fatigue), a disposition situation (eg, admission/discharge from the emergency department), a drug class (eg, antihypertensive agents, antibiotics).
2. **Write the lead-in for the set** (eg, *For each patient described below, select the most likely diagnosis*). The lead-in indicates the relationship between the stems and options, clarifying the question posed for examinees. It is an essential component of an Extended-Matching set.
3. **Prepare the list of options.** The list of options should be single words or very short phrases. List the options in alphabetical order unless there is a logical order.
4. **Write the items.** The items within a set should be similar in structure. Most often, patient vignettes are appropriate.
5. **Review the items.** Check to make sure that there is only a single “best” answer for each question. Also make sure that there are at least four reasonable distractors for each item. As a final check, it is recommended that you ask a colleague to review the items (without the correct answer indicated). If the colleague has difficulty determining the correct answer, modify the option list or the item to eliminate the ambiguity.

Additional information on writing Extended-Matching items can be found in:

Case SM, Swanson DB. Extended-matching items: a practical alternative to free-response questions. *Teaching and Learning in Medicine*. 1993;5(2):107-115.

Case SM, Swanson DB, Woolliscroft JO. Assessment of diagnostic pattern recognition skills in medicine clerkships using a written test. In: Harden R, Hart I, Mulholland H, eds. *Approaches to Assessment of Clinical Competence*. Norwich, England: Page Brothers; 1992:452-458.

Sample Extended-Matching Sets

Sample Anatomy Set

- | | |
|-----------------------------------|------------------------------------|
| A. Left anterior cerebral artery | E. Left posterior cerebral artery |
| B. Right anterior cerebral artery | F. Right posterior cerebral artery |
| C. Left middle cerebral artery | G. Left lenticulostriate arteries |
| D. Right middle cerebral artery | H. Right lenticulostriate arteries |

For each patient with neurologic abnormalities, select the artery that is most likely to be involved.

1. A 72-year-old right-handed man has weakness and hyperreflexia of the right lower limb, an extensor plantar response on the right, normal strength of the right arm, and normal facial movements. **Ans: A**
2. A 68-year-old right-handed man has right spastic hemiparesis, an extensor plantar response on the right, and paralysis of the lower two-thirds of his face on the right. His speech is fluent, and he has normal comprehension of verbal and written commands. **Ans: G**

Sample Pharmacology Set

- | | |
|-------------------------|-------------------|
| A. Acetaminophen | J. Nalidixic acid |
| B. Amiodarone | K. Nitrofurantoin |
| C. ACE inhibitors | L. Penicillin |
| D. Aspirin | M. Prednisone |
| E. Atenolol | N. Procainamide |
| F. Bleomycin | O. Propranolol |
| G. Cytosine arabinoside | P. Sulfasalazine |
| H. Furosemide | Q. Tetracycline |
| I. Metronidazole | R. Verapamil |

For each patient, select the drug most likely to have caused the adverse effect.

1. A 56-year-old man with recurrent ventricular arrhythmias began taking an antiarrhythmic drug 5 months ago. He now has progressive dyspnea, cough, and low-grade fever. Erythrocyte sedimentation rate is increased. X-ray film of the chest shows a diffuse interstitial pneumonia. Pulmonary function tests show that diffusing capacity for carbon monoxide is decreased. **Ans: B**
2. A 62-year-old man with chronic obstructive pulmonary disease begins therapy with an antihypertensive drug. Two weeks later, he has marked worsening of dyspnea and clearly audible wheezing. **Ans: O**

Sample Physiology Set — Clinical Features (Which additional findings are likely?)

	<u>pH</u>	<u>PO₂ mm Hg</u>	<u>PCO₂ mm Hg</u>	<u>HCO₃ - mEq/L</u>
A.	7.15	98	33	11
B.	7.15	98	24	8
C.	7.30	56	80	38
D.	7.40	100	40	25
E.	7.50	100	33	25
F.	7.50	100	24	18
G.	7.50	56	33	25

For each patient described below, select the most likely arterial blood gas findings.

1. A 22-year-old man with a 3-week history of polyuria and polydipsia has had nausea, vomiting, and decreased responsiveness for the past 12 hours. Urinalysis (dipstick) shows 4+ glucose and 4+ ketones. **Ans: B**
2. A 25-year-old woman is brought to the emergency department 12 hours after a suicide attempt. She took approximately 100 500-mg aspirin tablets. **Ans: F**

Sample Diagnosis Set

- | | |
|----------------------------------|------------------------------|
| A. Ankylosing spondylitis | E. Osteoporosis |
| B. Intervertebral disc infection | F. Spinal stenosis |
| C. Multiple myeloma | G. Spondylolysis |
| D. Myofascial pain | H. Tuberculosis of the spine |

For each patient with back pain, select the most likely diagnosis.

1. A 26-year-old man has insidious onset of low back pain and early morning stiffness. The pain alternates from side to side and occasionally radiates into the buttocks and back of the thighs, but not below the knees. The patient has acute anterior uveitis, diffuse low back and sacroiliac tenderness, and restricted range of motion at the hips. His erythrocyte sedimentation rate is 40 mm/h; latex fixation test is negative; and mild hypoproliferative anemia is present.

Ans: A

2. Twelve hours after being struck from the rear while driving her car, a 28-year-old woman has vague cervical and lumbar pain associated with headache and restricted cervical mobility. She is now very anxious. Rope-like bands of muscle are present in the lumbar area and over the left buttock; the bands are painful

Ans: D

Additional items would cover some of the remaining diagnoses. The sample vignettes are of average length; shorter, more focused vignettes may also be used. Alternatively, examinees could be challenged to identify key diagnostic information by using longer vignettes.

Sample Diagnosis Set

- | | |
|---------------------------------|---------------------------------|
| A. Abdominal aneurysm | K. Kidney stone |
| B. Appendicitis | L. Mesenteric adenitis |
| C. Bowel obstruction | M. Mesenteric artery thrombosis |
| D. Cholecystitis | N. Ovarian cyst — ruptured |
| E. Colon cancer | O. Pancreatitis |
| F. Constipation | P. Pelvic inflammatory disease |
| G. Diverticulitis | Q. Peptic ulcer disease |
| H. Ectopic pregnancy — ruptured | R. Perforated peptic ulcer |
| I. Endometriosis | S. Pyelonephritis |
| J. Hernia | T. Torsion |

For each patient with abdominal pain, select the most likely diagnosis.

1. A 25-year-old woman has sudden onset of persistent right lower abdominal pain that is increasing in severity. She has nausea without vomiting. She had a normal bowel movement just before onset of pain. Examination shows exquisite deep tenderness to palpation in right lower abdomen with guarding but no rebound; bowel sounds are present. Pelvic examination shows a 7-cm, exquisitely tender right-sided mass. Hematocrit is 32%. Leukocyte count is 18,000/mm³. Serum amylase activity is within normal limits. Test of the stool for occult blood is negative.

Ans: B

2. An 84-year-old man in a nursing home has increasing poorly localized lower abdominal pain recurring every 3-4 hours over the past 3 days. He has no nausea or vomiting; the last bowel movement was not recorded. Examination shows a soft abdomen with a palpable, slightly tender, lower left abdominal mass. Hematocrit is 28%. Leukocyte count is 10,000/mm³. Serum amylase activity is within normal limits. Test of the stool for occult blood is positive.

Ans: E

Sample Management Set: Disposition

- A. Observe in emergency department
- B. Admit for surgery
- C. Admit for medical management
- D. Admit for endoscopy
- E. Admit for laparoscopy
- F. Order contrast studies
- G. Order MRI
- H. Order CT scan
- I. Order ultrasonography
- J. Send home with analgesics
- K. Send home for follow-up by personal physician

For each of the following patients, select the most appropriate next step in patient care.

Items might describe patients with appendicitis, ectopic pregnancy, endometriosis, Crohn's disease, diverticulitis, pelvic abscess, sickle cell crisis, renal lithiasis, twisted ovarian cyst, or other problems that commonly present as emergencies. Other disposition sets might focus on "telephone triage," hospital transfer/discharge decisions, etc.

Sample Management Set: Diagnostic Testing

- | | |
|---------------------------------------|---------------------------------|
| A. Test of the stool for occult blood | H. Exercise tolerance test |
| B. Fasting serum glucose level | I. Digital prostate examination |
| C. Hemoglobin level | J. ECG |
| D. Prostate-specific antigen level | K. Spirometry |
| E. Serum cholesterol level | L. X-ray film of the chest |
| F. Serum iron level | M. Sigmoidoscopy |
| G. Thyroid function tests | |

For each patient who comes to the physician for a health maintenance examination, select the most appropriate diagnostic study.

1. A 22-year-old man who weighs 89 kg (196 lb) and is 175 cm (69 in) tall has smoked one pack of cigarettes daily for 8 years; he does not exercise. His last examination was 5 years ago. His father had a myocardial infarction at the age of 48 years. Physical examination shows no abnormalities. **Ans: E**
2. A 28-year-old woman who weighs 70 kg (154 lb) and is 173 cm (68 in) tall has smoked one pack of cigarettes daily for 12 years; she does not exercise. Her last examination was 5 years ago, though she had a Pap smear 9 months ago that showed normal results. Her father had a myocardial infarction at the age of 48 years. Her grandmother was diagnosed with colon cancer at the age of 62 years. Physical examination shows no abnormalities. **Ans: E**

Sample Option List for Electrolyte Abnormalities

- | | |
|-------------------|--------------------|
| A. Hypocalcemia | E. Hypercalcemia |
| B. Hypokalemia | F. Hyperkalemia |
| C. Hypomagnesemia | G. Hypermagnesemia |
| D. Hyponatremia | H. Hybernatriemia |

For each of the following patients, select the electrolyte abnormality most likely to be present.

Sample Behavioral Sciences/Pediatrics Option List

	Cognitive/ Language Skills	Gross Motor Skills	Social Skills
A.	Normal	Normal	Normal
B.	Normal	Normal	Delayed
C.	Normal	Delayed	Normal
D.	Normal	Delayed	Delayed
E.	Delayed	Normal	Normal
F.	Delayed	Normal	Delayed
G.	Delayed	Delayed	Normal
H.	Delayed	Delayed	Delayed

For each child, select the best description of development.

Sample Nutrition/Biochemistry Option List

A.	Vitamin A	I.	Folate
B.	Vitamin B ₁ (thiamine)	J.	Biotin
C.	Vitamin B ₂ (riboflavin)	K.	Niacin
D.	Vitamin B ₆	L.	Iron
E.	Vitamin C	M.	Magnesium
F.	Vitamin D	N.	Copper
G.	Vitamin E	O.	Zinc
H.	Vitamin K	P.	Iodine

For each child with a metabolic abnormality, select the vitamin or mineral that is most likely to be involved.

See Appendix B for additional examples.

Steps for Organizing a Group to Write Clinical R-Sets

The following steps may be followed to utilize a group in writing clinical R-sets. Some groups have met over dinner; followed the steps below; and generated a first draft of a dozen or more items from each attendee. Others have scheduled this as a full-day off-campus retreat, with the goal of generating a pool of near-final items.

The organizer of the “item writing party” should think about how the participant’s time can best be spent. The yield will be substantially greater in terms of both quality and quantity if some work is done in advance. For example, there will be significantly smaller yield if the participants are expected to decide what topics to write on; time will be saved if the topics for each set are defined in advance of the meeting (Step #1 below). Similarly, more items will be developed if a draft of the options for each set is developed in advance (Step #3 below). More items will also be developed if a sample item is written as a model for each set (Step #4 below). There are situations where it is best to allow item writers considerable flexibility in determining what to write. In these circumstances, providing one or more option sets and allowing them to generate one or more options sets on their own might be workable.

Decisions will also need to be made about the composition of the pairs of item writers. In some circumstances, it might be best to allow participants to self-select their partners. In other circumstances, it might be best to assign the pairs. We have had the most success assembling item-writing pairs with similar interests, but including individuals with more diverse expertise as reviewers. If the goal is to generate interdisciplinary items for an exam, the same list of options may be provided to several pairs of item writers. For example, options related to abdominal pain could be given to a pair of gynecologists; a pair of surgeons; and a pair of internists. In constructing the exam, one or more items from each discipline could be merged into the same set, requiring that examinees think across disciplines in determining the most likely diagnoses of the patients.

Significant time can be saved if participants write their items on computers. This saves many hours of trying to decipher the handwriting of the item writers. We typically have two item writers work at one computer (we have found that someone automatically assumes control of the keyboard). Then, several approaches are possible for the review: reviewers may congregate around the computer to read the items on screen; the item author can read the item aloud to the reviewers who do not have a copy of the items; or hard copies can be printed for everyone to use in the review session.

- 1. Define the content domain of the exam.** For example, in developing an exam to test the ability to diagnose common clinical problems, you might define the domain by a list of the chief complaints. Time will be saved if this is done in advance of the item-writing meeting.

Abdominal mass	Dizziness	Lymphadenopathy
Abdominal pain	Easy bruising	Movement abnormalities
Anemia	Fatigue	Nausea
Ascites	Fever	Palpitations
Back pain	Gastrointestinal bleeding	Sexual dysfunction
Chest pain	Headache	Shortness of breath
Confusion	Hematemesis/Melena	Skin lesion
Cough	Itching	Syncope
Developmental delay	Jaundice	Weight change
Diarrhea	Joint pain	Wheezing

- 2. Train a group of faculty members to serve as item writers.** Training should include a brief discussion of the purpose of the exam, some sample items, and the procedures to be followed during item writing.
- 3. Divide the group into pairs to write items.** Each pair is assigned to write on 2-4 chief complaints; they generate (or modify) a list of diagnoses for each assigned complaint and write one or more patient descriptions for the diagnoses they included in their option list. Expect 20 to 60 item stems from each pair (10 to 20 per complaint). Use of computers will save considerable time in the long run.

4. Stress the following guidelines for writing stems.

Each item should describe a patient with one of the diagnoses in the option list, beginning with the patient's age, gender, chief complaint, and site of care, followed by personal history, family history (if relevant), then physical examination information, then laboratory data (if provided).

Depending upon the purpose of the set, vignettes can be brief prototypic presentations or fuller descriptions that challenge examinees to identify key information.

Each patient description should be similar in structure to the others in the set. For example, if race, ethnicity, or occupation is included in one item, include it in all items; if laboratory data are included for one item, include them in all items.

5. Merge the pairs into a larger group to review the items. One approach is to have the author read the item aloud; others attempt to provide the correct answer. The group reviews the option list and modifies the item or the option list to eliminate any ambiguity. Other approaches are outlined above.

6. Type, edit, and subject the items to external review. Items should be reviewed without the correct answer indicated after they are in their final form.

7. Construct the test. Select a sample of items from each complaint; save the remaining items for subsequent exams. Items can be converted into one-best-answer items by adding a lead-in and the best five (or more) options from the option list.

Form for Writing R-Sets

Theme: _____
(eg, a presenting complaint)

Lead-In: _____
(eg, For each patient with fever, select the most likely diagnosis.)

Options

A	N
B	O
C	P
D	Q
E	R
F	S
G	T
H	U
I	V
J	W
K	X
L	Y
M	Z

(Write items on separate pages)

Sample SPSSX Code to Score Multiple-Choice Tests, including Extended-Matching Items

The following SPSSX code can be used as a model for scoring a hypothetical test including up to 100 multiple-choice items; each item may have up to 26 options but only one correct answer. It is straightforward to alter the code for any test length.

It is assumed that:

- the answer key is in a file named KEY.DAT (format: an eight-character exam ID, followed by a space and the 100 correct answers);
- the examinee responses are in a file named RESPONSE.DAT (format: social security number or any nine-digit examinee ID, a space, the eight-character exam ID, a space, the examinee's responses to the 100 items);
- the item analysis output shows the distribution of responses to each item, plus a reliability coefficient (coefficient alpha); and
- score reports with the examinee ID number, plus percent correct and standard scores (placed in a file named REPORT.LIS). REPORT.LIS can be imported into a word processor to enhance its appearance.

For experienced SPSSX users, the code should be easy to understand. SPSSX provides some file management facilities that vaguely resemble those in relational data bases (illustrated in the AGGREGATE and MATCH FILES commands); these join the answer key to each examinee record, calculate a mean and SP, and join these to each examinee record.

It is straightforward to generalize the code to handle subscores, more complicated scoring algorithms, etc.

```

TITLE          SAMPLE SPSSX SETUP FOR SCORING A HYPOTHETICAL 100-ITEM
SUBTITLE      MULTIPLE CHOICE TEST — DAVE SWANSON, SEPTEMBER 6, 1991
SET           LENGTH=64/WIDTH=132

COMMENT      READ IN AND SAVE THE ANSWER KEY
FILE HANDLE  KEYDAT/NAME='KEY.DAT'
DATA LIST    FILE=KEYDAT/
             EXAMCODE,KEY1 TO KEY100
             (A8,1X,100A1)

FILE HANDLE  KEYSYS/NAME='KEY.SYS'
SAVE        OUTFILE=KEYSYS

COMMENT      READ IN EXAMINEE RESPONSE STRINGS
FILE HANDLE  RESPDAT/NAME='RESPONSE.DAT'
DATA LIST    FILE=RESP/
             SSN,EXAMCODE,RESP1 TO RESP100
             (F9.0,1X,A8,1X,100A1)

COMMENT      ADD THE ANSWER KEY TO RESPONSE STRING RECORDS
MATCH FILES  FILE=*/TABLE=KEYSYS/BY EXAMCODE

COMMENT      COMPARE THE KEY TO RESPONSES AND CREATE A 0/1 VECTOR OF
COMMENT      INCORRECT/CORRECT ANSWERS
VECTOR       SCORE(100,F1.0)
DO REPEAT    K=KEY1 TO KEY100/R=RESP1 TO RESP100/S=SCORE1 TO SCORE100
COMPUTE      S=0
IF           (K EQ R) S=1
END REPEAT

COMMENT      PRINT A CROSSTABULATION OF RESPONSES FOR EACH ITEM
TABLES       FORMAT=CWIDTH(10,3) NSPACE LIGHT/
             TABLE= RESP1 +
             RESP2 +
             ... + (you'd actually need to type all of these in)
             RESP100 BY (LABELS)/
             STATISTICS=COUNT((F3.0) ``)

COMMENT      GENERATE RELIABILITY STATISTICS
RELIABILITY  VARIABLES=SCORE1 TO SCORE100/
             SCALE(TOTAL)=SCORE1 TO SCORE100/
             STATISTICS=DESCRIPTIVE,SCALE,ANOVA/
             SUMMARY=ALL

COMMENT      CALCULATE A PERCENT CORRECT SCORE FOR THE TOTAL TEST
COMPUTE      PCSCORE=100*MEAN(SCORE1 TO SCORE100)

COMMENT      CALCULATE A STANDARD SCORE FOR THE TOTAL TEST
FILE HANDLE  MEANSD/NAME='MEANSD.TMP'
AGGREGATE    OUTFILE=MEANSD/BREAK=EXAMCODE/
             PCMEAN=MEAN(PCSCORE)/PCSD=SD(PCSCORE)
MATCH FILES  FILE=*/TABLE=MEANSD/BY EXAMCODE
COMPUTE      STDSCORE=500+100*(PCSCORE-PCMEAN)/PCSD

COMMENT      WRITE OUT A SCORE REPORT
FILE HANDLE  REPORT/NAME='REPORT.LIS'
WRITE        OUTFILE=REPORT/
             SSN,PCSCORE,STDSCORE
             (3F9.0)

EXECUTE
FINISH

```

Comparison of Items in Five-Option and Extended-Matching Format

In several studies to investigate the optimal number of options for multiple-choice items, we have consistently found that, other things being equal, more options are better than fewer options. Based on items used on NBME exams, extended-matching items are more discriminating than any other format; 5-option A-types are second best; and the various forms of true/false items are the worst. In controlled studies comparing otherwise equivalent 5-option and extended-matching items, extended-matching items were found to be more discriminating than 5-option items; comparable levels of reproducibility can be achieved with the extended-matching format using one-third fewer items than with 5-option items. Extended-matching items were also found to be more difficult than content-parallel 5-option items: there is a lower probability of guessing the correct answer, and item writers are not always able to select the most functional distractors in reducing the number of options to five.

The following table shows examinee responses to a sample item presented in a 5-option format and in a 15-option format. The item was originally written as a 15-option item; the item writer then reduced the number of options to five by selecting what he thought were the best distractors (B, F, G, J, N). The item was markedly easier in the 5-option format (p value of 81 vs 59), and the discrimination was markedly lower (not shown). There is an increased probability of examinees selecting the correct answer in the 5-option format, especially because item writers do not uniformly identify the most salient distractors (eg, option D). Tests constructed of extended-matching items tend to spread out the lower ability students; the extended list of options gives them more opportunity to show what they don't know.

Examinee Responses

Format	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5-option		81		*		4	0			14			*	1	
15-option	2	59	2	13	1	2	1	0	1	7	2	2	8	0	0

For additional information, see also:

Case SM, Swanson DB, Ripkey DR. Comparison of items in five-option and extended-matching format for assessment of diagnostic skills. *Academic Medicine*. 1994;69(suppl):S1-S3.

Case SM, Swanson DB. Extended-matching items: a practical alternative to free-response questions. *Teaching and Learning in Medicine*. 1993;5(2):107-115.

Swanson DB, Case SM. Trends in written assessment: a strangely biased perspective. In: Harden R, Hart I, Mulholland H, eds. *Approaches to Assessment of Clinical Competence*. Norwich, England: Page Brothers; 1992:38-53.

A's to R's and Back Again

Although research has consistently shown a psychometric advantage of Extended-Matching (R-type) items over five-option A-type items, there may be circumstances where you need to convert items from one format to another. This should be straightforward.

For example, the following item was written as a five-option A-type item:

A patient with the classic phenotypic features of trisomy 21 (Down syndrome) has 46 chromosomes on each of 100 metaphase karyotypes. Which of the following is the most likely explanation for this finding?

- | | |
|----------------------------|------------------------------|
| <i>A. Deletion</i> | <i>*D. Translocation</i> |
| <i>B. Mosaicism</i> | <i>E. Undetected trisomy</i> |
| <i>C. Somatic mutation</i> | |

This can easily be transformed into the following extended-matching item. Once in this format, additional stems can be written to expand the set.

- | | |
|----------------------------|------------------------------|
| <i>A. Deletion</i> | <i>D. Translocation</i> |
| <i>B. Mosaicism</i> | <i>E. Undetected trisomy</i> |
| <i>C. Somatic mutation</i> | |

For each patient with genetic abnormalities, select the genetic pattern that is most likely to be involved.

A patient with the classical phenotypic features of trisomy 21 (Down syndrome) has 46 chromosomes on each of 100 metaphase karyotypes.

Additional options could be written to make the examinee task more challenging.

- | | |
|------------------------------|---------------------------------|
| <i>A. Deletion</i> | <i>F. Somatic mutation</i> |
| <i>B. Genomic imprinting</i> | <i>G. Translocation</i> |
| <i>C. Mosaicism</i> | <i>H. Undetected trisomy</i> |
| <i>D. Pleiotropy</i> | <i>I. Variable expressivity</i> |
| <i>E. Reduced penetrance</i> | |

For each patient with genetic abnormalities, select the genetic pattern that is most likely to be involved.

Items from this set could be converted back to single A-type items with five or more options.

Chapter 7

Pick N Items: An Extension of the Extended-Matching Format



The Pick N format may be similar to either the extended-matching or the A-type format; the primary difference is that the examinee is told to pick 2, 3, 4, or even 5 of the options listed. As with extended-matching sets, the option list may include up to 26 options. The format was developed to replace negative items or items with double options, particularly in areas such as Health Maintenance and Disease Prevention. Items might focus on various patients with different risk factors who come for routine examination; the examinee would be asked to select the laboratory studies or immunizations that should be ordered for each patient. The format could also be used for emergency management items where several procedures would be carried out simultaneously; the examinee would be asked to select a specific number of actions from the option list.

The item-writing rules are the same as for extended-matching sets. The options should be short (usually a single word or very short phrase); the patient vignettes can be long. If the set asks for management decisions, each vignette should contain all relevant history and physical examination data. As with extended-matching items, the format works well for items that appear to be extremely easy; tricky or unnecessarily complex vignettes should be avoided.

The Pick N format is designed to specify exactly how many options to select. The rationale for this decision is derived from the essential difference between true/false and one-best-answer items, where true/false items require the examinee to indicate all responses that are appropriate, and one-best-answer items require the examinee to indicate a specific number of responses. Specifying the number of options to be selected changes the task from a true/false task to a best-answer task.

Research indicates that partial credit scoring is preferred but logistical considerations might preclude this. If you use all-or-nothing scoring, the items may be extremely difficult, and it is best to require examinees to select only two or three options, rather than more.

For additional information, see also:

Ripkey DR, Case SM, Swanson DB. A “new” item format for assessing aspects of clinical competence. *Academic Medicine*. 1996;71 (suppl):S34-S36.

- | | |
|-------------------------------|------------------------------|
| A. Diabetic polyneuropathy | F. Multiple sclerosis |
| B. Huntington's disease | G. Parkinson's disease |
| C. Lateral medullary syndrome | H. Pontine glioma |
| D. Lead encephalopathy | I. Tabes dorsalis |
| E. Medulloblastoma | J. Wernicke's encephalopathy |

A 50-year-old man has the gradual onset of mental confusion, disorientation, and loss of short-term memory. He has a left footdrop. A blood smear shows microcytosis and basophilic stippling of erythrocytes. (Select the two most likely diagnoses)

In the example above, there would be disagreement about what diagnoses are likely, but the task becomes clear if the examinee is told to select the two most likely diagnoses. The options can be diagnosed as follows.

F	C	E	D	A	G	H	B
<i>Least Likely Diagnosis</i>						<i>Most Likely Diagnosis</i>	

Sample Pick N Set

- | | |
|--------------------------------------|---|
| A. Calcium | G. Vitamin B ₆ |
| B. Fluoride | H. Vitamin B ₁₂ (cyanocobalamin) |
| C. Folic acid | I. Vitamin C |
| D. Iron | J. Vitamin D |
| E. Vitamin A | K. Vitamin E |
| F. Vitamin B ₁ (thiamine) | |

For each child, select the appropriate vitamin or mineral supplements.

1. A 1-month-old infant is brought to the physician for a well-child examination. He has been exclusively breast-fed, and examination shows normal findings. (SELECT 2 SUPPLEMENTS). **Ans: B, J**
2. A 6-year-old girl has cystic fibrosis. She has been taking no medications. (SELECT 3 SUPPLEMENTS). **Ans: E, J, K**

Sample Pick N Set

- | | |
|--|------------------------------|
| A. Analysis and culture of cerebrospinal fluid | F. Urinalysis |
| B. Blood culture | G. Urine culture |
| C. Complete blood count | H. X-ray film of the abdomen |
| D. Examination of the stool for leukocytes | I. X-ray film of the chest |
| E. Measurement of serum electrolyte levels | |

For each child with fever, select the appropriate initial diagnostic studies.

1. A previously healthy 1-year-old girl is brought to the emergency department because of fever for 1 day. Her temperature is 41 C (105.8 F). She is otherwise asymptomatic. Physical examination shows no abnormalities. (SELECT 4 STUDIES). **Ans: B, C, G, I**
2. A previously healthy 10-day-old newborn is brought to the emergency department because of fever for 2 hours. He was born at term after an uncomplicated pregnancy. His temperature is 39 C (102.2 F). Physical examination shows no abnormalities. (SELECT 6 STUDIES). **Ans: A, B, C, E, G, I**
3. A 7-year-old boy with sickle cell disease is brought to the emergency department because of fever for 1 day and chest pain for 1 hour. His temperature is 39.5 C (103.1 F). Breath sounds are slightly decreased in the right lower lung; he is not in respiratory distress. (SELECT 3 STUDIES). **Ans: B, C, I**

PATIENT CHART**Patient History**

Sex: male
Current age: 28 years
Chief complaint: health maintenance examination
Social history:
Marital status: single
Occupation: computer programmer
Alcohol: 2-4 beers/weekend
Smoking: 5-10 cigarettes daily from age 16-24
Exercise prgm: sedentary
Medical history:
Childhood: obese since grade school
Immunizations: all childhood immunizations; last tetanus toxoid age 15;
no immunizations since childhood
Screening: no physician visits since college
Family history:
Parents: father age 57; hypertensive mother age 55; obese and hypertensive
Siblings: none
Children: none
Current medications: none
Allergies: none

Physical Examination

Height: 178 cm (70 in)
Weight: 134 kg (295 lb)
Vital signs:
Blood pressure 148/86 mm Hg
Pulse 90/min
Respirations 16/min
Skin: erythematous rash in groin
Abdominal: obese
Laboratory studies: none ordered

1. For the patient whose chart is shown, select the conditions for which he is at increased risk.
(SELECT 4 CONDITIONS)

- | | |
|----------------------------|----------------------------|
| A. Alcoholism | G. Hypertension |
| B. Colon cancer | H. Hypothyroidism |
| C. Coronary artery disease | I. Osteoarthritis |
| D. Diabetes mellitus | J. Skin cancer |
| E. Hemochromatosis | K. Thyroid cancer |
| F. HIV | L. Urinary tract infection |

2. For the patient whose chart is shown, select the most appropriate health maintenance interventions.
(SELECT 4 INTERVENTIONS)

- | | |
|--------------------------------|--|
| A. Complete blood count | H. Serum urea nitrogen (BUN) and creatinine level measurements |
| B. Dietetic counseling | I. Tetanus toxoid vaccine |
| C. Exercise counseling | J. Thyroid function tests |
| D. Fasting serum lipid profile | K. Urinalysis |
| E. Hepatitis B immunization | L. X-ray film of the chest |
| F. HIV testing | |
| G. Influenza immunization | |

Section IV

Additional Issues



This section includes some additional issues that are related to testing.

Chapter 8

Interpretation of Item Analysis Results



Many schools provide faculty with item analysis output following each multiple-choice examination. This output is an excellent source of information about an item and is useful in evaluating the quality of the item, as well as in evaluating the accuracy of the answer key.

The following are sample results from four items; each illustrates a common situation. The students taking the test were divided into a Hi group and a Lo group, based on their performance on the total test. If you have a small number of examinees, include the top 50% of the students in the Hi group and the bottom 50% in the Lo group. If you have a large number of examinees, you might include the top 25% in the Hi group and the bottom 25% in the Lo group.

Typically, item analysis output indicates the percentage of students in each group who selected each option. Often, it also includes some measure of item difficulty (eg, the “p-value” or the proportion of students who answered the item correctly) and some measure of discrimination (eg, a biserial or a point biserial). We recommend that attention be focused on the pattern of responses rather than on the difficulty level or discrimination index.

For each sample item below, the percentage of students selecting each option is shown. The total row shows the percentage of the total group who selected each option. For example, in Item #1, 1% of the Hi group selected Option A; 1% selected B; 91% selected C; 4% selected D; 1% selected E; and 2% selected F. In the same item, 20% of the Lo group selected Option A; 6% selected B, etc. The asterisk on Option B indicates that B was the purported correct answer.

Item #1

Group	A	B*	C	D	E	F
Hi	1	1	91	4	1	2
Lo	20	6	51	14	6	3
Total	9	2	76	8	3	2

p-value: 2 discrimination index: -0.21

Interpretation: This is the typical pattern for an item that is miskeyed: if the answer is Option B, the item is very difficult and the discrimination index is negative. With a key of B, only 2% of the students answered correctly. The correct answer is almost certainly C, but a content expert should review the item to make sure. If the correct answer is C, the p-value becomes 76 and the discrimination index becomes 0.46 — these are both excellent from a statistical perspective, and there is no reason to make any changes in item text.

Item #2

Group	A	B	C*	D	E	F
Hi	0	1	90	3	3	3
Lo	0	1	60	25	8	6
Total	0	1	74	12	7	6

p-value: 74 discrimination index: 0.33

Interpretation: 90% of the Hi group and 60% of the Lo group selected the correct answer. These are excellent overall statistics. You could rewrite A and B before you reuse the item because few students selected those options.

Item #3

Group	A	B	C*	D	E	F
Hi	44	1	50	2	1	2
Lo	20	15	21	22	20	2
Total	32	7	34	14	11	2

p-value: 34 discrimination index: 0.30

Interpretation: 50% of the Hi group and 21% of the bottom group selected the correct answer. This is a very difficult item that is probably NOT OK. Too many of the Hi group selected Option A; the item may be poorly worded. Check Option A for “fairness.” Make sure Option A is not equally correct.

Item #4

Group	A	B	C*	D	E	F
Hi	18	10	51	17	2	2
Lo	24	24	21	25	4	2
Total	22	17	34	22	3	2

p-value: 34 discrimination index: 0.30

Interpretation: The Hi/Lo group breakdown on option C is identical to Item #3, but this item may be OK. In contrast to Item #3, those who don't know the correct answer are widely spread across the various distractors. Of course, it would still be desirable to review options A, B and D for correctness and clarity.

Chapter 9

Establishing a Pass/Fail Standard

❖

Definitions and Basic Principles

Standards may be classified as either *relative* or *absolute*. A *relative standard* is based on the performance of the group taking the test. Examinees pass or fail depending upon how well they perform relative to other examinees taking the test. The following are examples of *relative* standards:

Those scoring below 1.2 standard deviations below the mean will fail.
The bottom 20 percent of the group will fail.

In contrast, an *absolute standard* does not compare the performance of one examinee with the others who are taking the test. Examinees pass or fail based only upon how well they perform, regardless of the performance of other examinees. All examinees could pass or all could fail. The following is an example of an *absolute* standard:

Those answering less than 60 percent of the questions correctly will fail.

Unless there are strong reasons to fail a given number of examinees, an absolute standard (based on examinee performance) is preferred over a relative standard (based on a particular failure rate).

Basic Principles of Setting Standards

- Regardless of the procedure used, setting standards requires judgement. Setting standards will always be arbitrary, but need not be capricious.
- Unless there is a specific reason to fail a given number of examinees (eg, there are only a fixed number of slots available), a standard based on examinee mastery of exam content is preferred over a standard based on a particular failure rate.

- It is wise to involve multiple informed judges in the standard-setting process. Differences of opinion will occur, and use of multiple judges will reduce hawk/dove effects.
- Judges should be provided with data on examinee performance at some point in setting standards. Setting standards without such data may lead to uninformed standards and unreasonable results.

A helpful “how-to” reference on standard setting is:

Livingston SA, Zieky MJ. *Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests*. Princeton, NJ: Educational Testing Service; 1982.

Two Standard-Setting Methods Based on Judgements about Items

The Modified Ebel Procedure

- A group discusses the characteristics of the “borderline examinee”: an examinee whose skills are just good enough to allow him/her to pass.
- Judges categorize items as “Essential,” “Important,” or “Indicated.”
- Judges indicate the number of items in each category that a borderline examinee would obtain.
- The pass/fail standard is calculated as the percentage of possible points that a borderline examinee would obtain.

The Modified Angoff Procedure

- A group discusses the characteristics of a “borderline” examinee.
- For each item on the test, the judges estimate the percentage of borderline examinees who would answer the item correctly.
- The pass/fail standard for the test is the average of the percentages for the items.

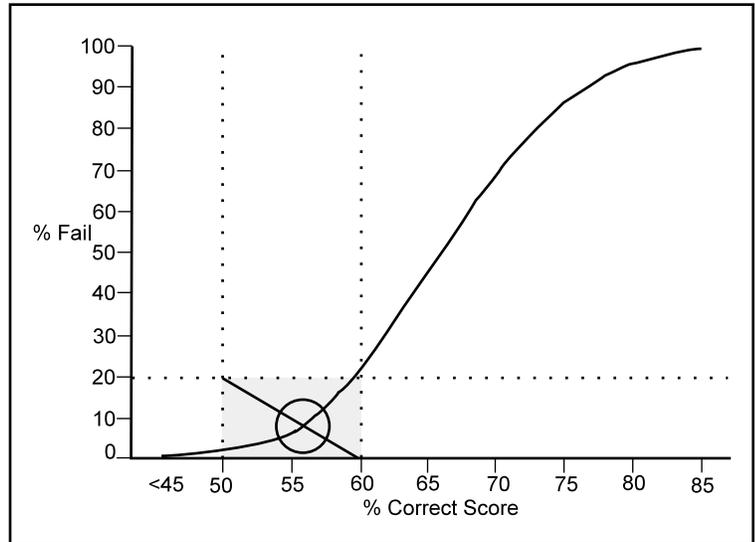
Common Variations on the Angoff Procedure

- Judges may or may not be provided with the correct answers to questions.
- Judges may or may not be provided with information concerning the percentage of examinees who answered each item correctly.
- After a period of training, judges may continue to work as a group or may work individually.

Relative/Absolute Compromise Standards: The Hofstee Method

More recently, several “compromise models” have been developed that utilize the advantages of both relative and absolute standard-setting procedures. One of these methods, the Hofstee method, is described below.

1. Judges are asked to review a copy of the exam.
2. Judges then indicate the following values, which define acceptable standards:
 - Lowest acceptable percentage of failing examinees (minimum failure rate)
 - Highest acceptable percentage of failing examinees (maximum failure rate)
 - Lowest score which would allow someone to pass (minimum passing point)
 - Highest score required for someone to pass (maximum passing point)
3. After test administration, a curve showing the fail rate as a function of passing score is plotted. (In the figure shown, the curve extends from bottom left to top right.)
4. The four values obtained in #2 are drawn, forming a rectangle. Often the median values of the group of judges are used. In the example, the appropriate failure rate was judged to be between 0 and 20% (see horizontal lines); the appropriate pass/fail point was judged to be between 50 and 60% correct (see vertical lines).
5. A line is drawn on the diagonal from upper left to lower right. The point where this intersects the curve is the standard (ie, just above 55% correct in the figure).



A useful reference on compromise methods is:

de Gruijter D. Compromise models for establishing examination standards. *Journal of Educational Measurement*. 1985;22:263-269.

Chapter 10

Miscellaneous Thoughts on Topics Related to Testing

❖

Comments on a hodge-podge of topics related to testing are provided below. In general, the points made are speculative and based on anecdotal experience rather than evidence. That is, they reflect our biases rather than the results of research.

Multiple Station Exams (a.k.a. Practical Exams, Steeplechases, OSCEs)

Though logistically complex to set up and administer, these are very useful in the basic sciences, particularly to assess hands-on skills that cannot be measured with paper-and-pencil tests (eg, ability to use a microscope, to perform a laboratory procedure). In addition, reproduction of some kinds of material (eg, results of imaging studies, color pictorial material) is very expensive; in such situations, the multiple-stations approach can be used to reduce test administration costs.

Take-Home Exams

Take-home exams can be a substantial learning experience for students by stimulating them to read broadly and deeply on important topics. Unfortunately, students tend to produce tomes as answers, and it can be unclear if submitted answers represent the student's own work. The same advantages can be gained by distributing (a superset of) test questions in advance, and administering (a subset of) questions as a timed test.

Open-Book Tests

Open-book tests can be a very good idea because of the impact on the kinds of questions that faculty prepare. For open-book tests, it is pointless to ask questions about isolated facts that can be looked up quickly on a single page of a text book, so test material developed for these tests tends to focus more on understanding of key concepts and principles in problem situations.

Frequent Short Quizzes versus Infrequent Tests

Infrequent testing makes each exam a major event; students may even stop attending class to prepare, and this seems undesirable. In addition, with infrequent tests, students may be unable to determine if they are studying the right material or learning in enough depth. Though it may be more time consuming for faculty, frequent testing reduces the importance of each individual exam and helps students to better gauge their progress. On the whole, frequent testing seems preferable, though students are likely to complain regardless of the approach adopted.

Keeping Tests “Secure” versus Permitting Students to Retain Them

Because tests can have a substantial “steering” effect on student learning, permitting students to retain test material can aid in focusing student attention on key topics, reinforcing curricular goals and course objectives (assuming test materials reflect these). However, preparation of good exam questions is very time consuming, and, over time, the quality of test material can deteriorate if faculty have to develop new test materials each time a course is taught. The best approach may be to make sample good-quality test material available in order to influence student learning, but maintain a bank of “secure” questions for repeated use, keeping in mind that security is likely to be poor, since students commonly memorize questions and reproduce them for each other.

Use of Cumulative Tests

Cumulative tests that hold students responsible for all material presented to date encourage attention to inter-relationships among topics, particularly if test questions require understanding of both old and newly presented topics. Use of tests that cover only material presented since the previous test encourages students to study topics in isolation; relationships among topics from different units may be missed. Since students can do badly on a series of tests because they never master basic material, this approach can also motivate students to remediate weaknesses.

Use of Integrative, Cross-Course Tests

Like the use of cumulative tests, integrative cross-course tests encourage students to see inter-relationships among disciplines and topics; this should be very helpful for long-term retention and for application of basic science knowledge to clinical situations. Generally, faculty from both basic science and clinical departments are needed for preparation of such exams. While time consuming, this joint effort may result in better test material as well as useful discussion among faculty of what material should be included in the curriculum.

Appendix A
The Graveyard of NBME Item Formats



Appendix A

The Graveyard of NBME Item Formats



The first National Board examination was administered in 1916. It was a five-day examination consisting of written, oral, laboratory, and clinical tests. Following World War I, the National Board appointed a commission to visit England, Scotland, and France to study methods of evaluation used for medical licensure. In 1922, the National Board administered a new examination: Part I was a three-day essay exam in the basic sciences; Part II was a two-day essay exam in the major clinical sciences; and Part III was a one-day oral exam conducted at the bedside. The examinations continued in those formats until the 1950s when multiple-choice testing became popular. In 1951, the National Board, with the cooperation of the Educational Testing Services, began a three-year study comparing the essay test with the multiple-choice test, and Part I and Part II were converted to a multiple-choice format in 1953; Part III was revised throughout the early 1960s as new formats were introduced and discontinued. The first exam consisted of primarily A-type items involving recall of facts, such as the following example:

In the 40 years since the first MCQ exam, the National Board has broadened the scope of the A-type item to test reasoning and problem-solving skills by including a clinical vignette in most item stems. Today, the A-type remains the most commonly used item format on the Step examinations. Many other item formats that were developed during this period have been discontinued. These formats (named by a letter in order of origination) are described on the following pages.

Of the following, the most effective prophylactic agent for the prevention of recurrences of rheumatic fever is

- A. *acetylsalicylic acid*
- B. *para-aminobenzoic acid*
- C. *adrenocorticotrophic hormone*
- D. *cortisone*
- E. *sulfadiazine*

The items shown in this section were obtained from:

Hubbard JP, Clemens WV. *Multiple-Choice Examinations in Medicine*. Philadelphia PA: Lea & Febiger;1961.

Critiques of these formats were obtained from:

Hubbard JP. *Measuring Medical Education*. Philadelphia PA: Lea & Febiger;1971.

Hubbard JP, Levit EJ. *The National Board of Medical Examiners: The First Seventy Years*. NBME;1985.

And also from various articles by Morton published in the *Federation Bulletin* in 1985 and 1986.

B-Type Items

B-type items were matching items that consisted of a list of lettered headings followed by a list of numbered words or phrases. The examinee was instructed to select the one heading that was most closely associated with each word or phrase. Because each response could be used more than once or not at all, B-type items could not be solved by elimination. B-type items were believed to widen the scope of an MCQ examination by allowing testing of a number of related subjects in a single series of items. Unlike the matching formats used today, the B-type items did not typically include a lead-in; as a result, the question being asked was sometimes unclear. These items generally performed well, and were only discontinued recently as the extended-matching format became widely used.

Sample B-type Item

DIRECTIONS: Each set of matching questions in this section consists of a list of three to five 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 and fill in the circle containing the corresponding letter on the answer sheet. Each lettered option may be selected once, more than once, or not at all.

- A. Coarctation of the aorta
 - B. Patent ductus arteriosus
 - C. Tetralogy of Fallot
 - D. Aortic vascular ring
 - E. Tricuspid atresia
-
- 1. Benefited by systemic-pulmonary artery anastomosis
 - 2. Most common type of congenital cyanotic heart disease
 - 3. Surgically corrected by resection and end-to-end anastomosis
 - 4. Possible cause of dysphagia in infants and children
 - 5. Hypertension in the arms and hypotension in the legs

D-Type Items

D-type items were complex matching items in which each item consisted of three functional disturbances (designated by a letter) and five situations (in a numbered list). The examinee was instructed to 1) select the functional disturbance or category that four of the five situations were related to and 2) indicate the one situation that did not belong in that category. 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 were confusing. In addition, they did not discriminate between knowledgeable and unknowledgeable examinees.

Sample D-type Item

DIRECTIONS: There are two responses to be made to *each* of the following questions. In the left-hand list are three lettered categories. Exactly four of the five numbered items in the right-hand list are related in some way to ONE of these categories. (1) on the appropriate line in the answer sheet blacken the space under the letter of the category in which these four items belong. (2) Then blacken the space under the number of the item in the right-hand list that does NOT belong in the same category with the other four.

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

K-Type Items

K-type items were the most commonly used multiple true/false item format at the National Board. They consisted of a stem followed by four options, one or more of which was correct. It was believed that K-type items tested in-depth knowledge or understanding of several aspects of a disease, a process, or a procedure, and required an examinee to be familiar with several different facts about a given topic. However, K-type items were criticized as being too complicated, requiring the examinee to constantly keep the answer code in mind. In addition, the possible response combinations introduced a cueing effect that reduced item discrimination and lowered test reliability. It was difficult to write good, unambiguous true/false items. 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"). K-type items were more difficult and less discriminating than other item types. In addition, they were less efficient than other MCQ formats, and the relative reliability per unit of test time was lower.

Sample K-type Item

<u>Directions Summarized</u>				
A	B	C	D	E
1, 2, 3 only	1, 3 only	2, 4 only	4 only	All are correct

1. A child suffering from an acute exacerbation of rheumatic fever usually has
 - (1) an elevated sedimentation rate
 - (2) a prolonged PR interval
 - (3) an elevated antistreptolysin O titer
 - (4) subcutaneous nodules

C-Type Items

C-type items were similar to B-type items in appearance but were multiple true/false in the task required of examinees. A C-type item consisted of a list of lettered headings followed by a list of numbered words or phrases. For each numbered item, the examinees were required to decide if A was true, B was true, both were true (option C), or neither was true (option D). This item type was used to compare and contrast two diseases, signs and symptoms, laboratory findings, etc. C-type items match K-types in level of difficulty. The primary problem with C-types was in deciding to what extent something had to be “true” to be selected. If, for example, something was associated with both A and B, but was more strongly associated with A, the examinee had to decide whether an appropriate response was A only or Both A and B. With relatively weak associations, the examinee had to decide whether the association was strong enough to note, or whether “neither” was the appropriate response. These judgements were not related to medical knowledge, but rather forced the examinee to think about what the item writers intended.

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 and fill in the circle containing the corresponding letter on the answer sheet. 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 ingestion of chloroquine once a week while in an endemic area.
3. Permanently cured by treatment with chloroquine.
4. Infection prevented by ingestion of chloroquine once a week.

E-Type Items

E-type items were multiple true/false items that are based on the analysis of relationships. Examinees who took E-type items still refer to them as the “True, True and Unrelated” items. The E-type consisted of a sentence with two main parts: an assertion and a reason for that assertion. The examinee was directed to select A if both were true statements and the reason was a correct explanation of the assertion; B if both were true statements but the reason was not a correct explanation of the assertion; C if the assertion was true but the reason was a false statement; D if the assertion was false but the reason was a true statement; E if both assertion and reason were false statements. It was thought that good reasoning skills and an understanding of the basic principles were required to answer this item type correctly. However, E-type items were difficult to construct, and examinees found them to be confusing.

Sample E-type Item

Directions Summarized			
A	True	True	Reason is a correct explanation.
B	True	True	Reason is NOT a correct explanation.
C	True	False	
D	False	True	
E	False	False	

Assertion

Reason

- | | | |
|--|---------|---|
| 1. Herpes simplex is usually regarded as an autogenous infection | BECAUSE | patients given fever therapy frequently develop herpes. |
| 2. Cow's milk is preferable to breast milk in infant feeding | BECAUSE | cow's milk has a higher content of calcium. |

H-Type Items

H-type items were comparison items that consisted of paired statements describing two entities to be compared in a quantitative sense. The examinee 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.

Although it was generally agreed that questions that depend on the memorization of absolute quantitative amounts should be limited, the H-type item was believed to be useful for those instances where recall of quantitative information was believed to be important. The problem for the examinees was in deciding 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. On the appropriate line of the answer sheet blacken the space under

A if (A) is greater than (B),

B if (B) is greater than (A),

C if the two are equal or very nearly equal.

- (A) The usual therapeutic dose of epinephrine
(B) The usual therapeutic dose of ephedrine
- (A) Life expectancy with glioblastoma of the occipital lobe
(B) Life expectancy with glioblastoma of the frontal lobe

I-Type Items

The I-type item was similar to the H-type. It consisted of pairs of phrases that describe conditions or quantities that might vary in relation to each other. The examinee was directed to select A if the two phrases were related directly (ie, an increase in the first was accompanied by an increase in the second or a decrease in the first was accompanied by a decrease in the second); B if the phrases were related inversely (ie, an increase in the first was accompanied by a decrease in the second or a decrease in the first was accompanied by an increase in the second); or C if the changes were independent of one another.

Sample I-type Item

DIRECTIONS: Each of the following pairs of phrases describe conditions or quantities that may or may not be related. On the appropriate line of the answer sheet blacken the space under

- 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

Neither the H- nor I-type formats were particularly popular. Because there were fewer options than in other item types, there was an increased chance of guessing the correct answer. In addition, the items tended to focus on minor details rather than scientific concepts.

In his series in the Federation Bulletin, Morton (1985-86) implied that different item types were included on medical licensure examinations simply to add variety to a lengthy examination. But, 25 years after the National Board converted from an essay exam to MCQ exams, the NBME reviewed the research on the various types of MCQs used, and the variety of item types was then reduced to include A-, B-, C-, G-, K-, X-, and M-type items. Staff again reviewed item types in the mid-1980s. The general consensus, at that time, was that four basic item types provided sufficient variety to test the knowledge specified as important for the awarding of a National Board certificate. These four basic types included A-, B-, C- and K-type items. G-types (sets of A-type items), N-types (sets of K-type items), and M-type items were no longer considered as separate formats.

More recently, the variety of item types has again been reviewed. The current Step examinations include A- and R-type items only. Some of the steps taken to improve the examinations include: concentrating on item types that are psychometrically sound, educating item writers on various item-writing techniques, focusing on clinical decision-making rather than recall items, and pretesting newly written items.

Appendix B
Sample Item-Writing Templates, Items,
Lead-Ins, and Option Lists
For the Basic and Clinical Sciences



Appendix B

Sample Item-Writing Templates, Items, Lead-Ins, and Option Lists for the Basic and Clinical Sciences



Gross Anatomy

Template for Gross Anatomy Items

Item Stem (Patient Vignette): (Describe a patient with a problem)

Lead-In: A defect is most likely to be present in which of the following structures/processes?

Options: (List of structures/processes)

A 65-year-old man has difficulty rising from a seated position and straightening his trunk, but he has no difficulty flexing his leg. Which of the following is most likely to be injured?

- A. * Gluteus maximus
- B. Gluteus minimus
- C. Hamstrings
- D. Iliopsoas
- E. Obturator internus

A 30-year-old man has loss of pain and temperature sensation on the left side of the face and on the right side of the body from the neck down; partial paralysis of the soft palate, larynx, and pharynx on the left; and ataxia on the left. This syndrome is most likely to result from thrombosis of which of the following arteries?

- A. Basilar
- B. *Right posterior inferior cerebellar
- C. Left posterior inferior cerebellar
- D. Right superior cerebellar
- E. Left superior cerebellar

Insert patient vignette describing a patient with a problem. Which of the following is the most likely disorder of the knee?

- A. Chondromalacia patellae
- B. Dislocation (tibiofemoral)
- C. Fracture of the patella
- D. Ganglion cyst
- E. Giant cell tumor of bone
- F. Osteochondritis dissecans
- G. Osteosarcoma
- H. Prepatellar bursitis
- I. Septic arthritis
- J. Tear of the meniscus

Insert patient vignette describing a patient with a problem. Which of the following is the vessel into which contrast medium should be injected during fluoroscopy to visualize the site of the abnormality?

- A. Celiac artery
- B. Internal iliac artery
- C. Inferior mesenteric artery
- D. Superior mesenteric artery
- E. Renal artery
- F. Portal vein

Insert patient vignette describing a patient with a problem. Which of the following gastrointestinal disorders is the most likely cause of these findings?

- A. Candidal esophagitis
- B. Diverticulitis
- C. Hiatal hernia
- D. Peptic ulcer
- E. Pseudomembranous enterocolitis
- F. Pyloric stenosis
- G. Regional enteritis
- H. Subphrenic abscess
- I. Ulcerative colitis

Isolated Fact Version:

Which of the following areas is supplied with blood by the posterior inferior cerebellar artery?

Application of Knowledge Version:

A 62-year-old man develops left-sided limb ataxia, Horner's syndrome, nystagmus, and loss of pain and temperature sensations on the face. Which of the following arteries is most likely to be occluded?

Behavioral Sciences

Insert patient vignette describing a child's age and what he can do. Which of the following best describes the level of development?

<u>Cognitive</u> <u>Language Skills</u>	<u>Gross Motor</u> <u>Skill</u>	<u>Social</u> <u>Skills</u>
A. Normal	Normal	Normal
B. Normal	Normal	Delayed
C. Normal	Delayed	Normal
D. Normal	Delayed	Delayed
E. Delayed	Normal	Normal
F. Delayed	Normal	Delayed
G. Delayed	Delayed	Normal
H. Delayed	Delayed	Delayed

A 55-year-old man comes to the emergency department because of chest pain. He is in no apparent distress. His temperature is 98.5°F, pulse is 68/min, respirations are 16/min, and blood pressure 130/74 mm Hg. Which of the following questions is most appropriate to ask first?

- A. Do you have a history of heart disease or myocardial infarction?
- B. Have you ever had chest pain before?
- C. How long did your chest pain last?
- D. Is the chest pain sharp or dull?
- E. Tell me about your chest pain?
- F. Where is your chest pain located?

A 35-year-old woman comes to the physician after finding a lump in her breast. She is crying and tells the physician that her mother died of breast cancer. Which of the following responses by the physician is the most appropriate?

- A. "I can see finding the lump was upsetting to you."
- B. "I doubt that this lump is breast cancer."
- C. "Tell me about how finding the lump made you feel."
- D. "Tell me more about the lump."
- E. "There is no reason to worry until we biopsy the lump."
- F. "Treatment for breast cancer has improved a lot since your mother died of breast cancer."

Isolated Fact Version:

Which of the following characterizes separation anxiety disorder?

Application of Knowledge Version:

An 8-year-old boy needs to be coaxed to go to school and, when at school, he often complains of severe headaches or stomach pain. Sometimes his mother has to take him home because of his symptoms. At night, he tries to sleep with his parents. When they insist that he sleep in his own room, he says that there are monsters in his closet. Which of the following best explains this behavior?

- A. Age-appropriate behavior
- B. Childhood schizophrenia
- C.* Separation anxiety disorder
- D. Social phobia
- E. Shared psychotic disorder

Biochemistry

Insert patient vignette describing a patient with a problem. Which of the following digestive enzymes or cofactors is most likely to be involved?

- A. Amylase
- B. Chymotrypsin
- C. Colipase
- D. Enterokinase
- E. Lactase
- F. Lipase
- G. Pepsin
- H. Sucrase
- I. Trypsin

Insert patient vignette describing a patient with a problem. Which of the following laboratory findings is the most likely?

- A. Hypercalcemia
- B. Hypocalcemia
- C. Hypermagnesemia
- D. Hypomagnesemia
- E. Hyponatremia
- F. Hyponatremia
- G. Hyperkalemia
- H. Hypokalemia

Isolated Fact Version:

Acute intermittent porphyria is the result of a defect in the biosynthetic pathway for

- A. collagen
- B. corticosteroid
- C. fatty acid
- D. glucose
- E. heme
- F. thyroxine

Application of Knowledge Version:

An otherwise healthy 33-year-old man has a 6-month history of episodes of mild weakness and steady, severe abdominal pain with some cramping but no diarrhea. One aunt and a cousin have had similar episodes. During an episode, his abdomen is distended, similar episodes. During an episode, his abdomen is distended, and bowel sounds are decreased. Neurologic examination shows mild weakness in the upper arms. These findings suggest a defect in the biosynthetic pathway for which of the following?

- A. Collagen
- B. Corticosteroid
- C. Fatty acid
- D. Glucose
- E.* Heme
- F. Thyroxine (T₄)

Isolated Fact Version:

Which of the following characterizes α -Fetoprotein?

Application of Knowledge Version:

A 45-year-old woman has rapid enlargement of the liver associated with deterioration of hepatic function associated with cirrhosis resulting from an episode of hepatitis. Serum concentration of which of the following is most likely to be abnormal?

- A. α -Antitrypsin
- B. Carcinoembryonic antigen
- C. Chorionic gonadotropin
- D.* α -Fetoprotein
- E. Gastrin

Non-Vignette Version:

An inherited metabolic disorder of carbohydrate metabolism is characterized by an abnormally increased concentration of hepatic glycogen with normal structure and no detectable increase in serum glucose concentration after oral administration of fructose. These two observations suggest that the disease is the result of the absence of which of the following enzymes?

- A. Fructokinase
- B. Glucokinase
- C.* Glucose 6-phosphatase
- D. Phosphoglucomutase
- E. UDPG-glycogen transglucosylase

Vignette Version:

A 6-month-old infant has an enlarged liver. Evaluation for metabolic disease shows an abnormally increased concentration of hepatic glycogen with normal structure and no detectable increase in serum glucose concentration after oral administration of fructose. These two observations suggest that the disease is a result of the absence of which of the following enzymes?

- A. Fructokinase
- B. Glucokinase
- C.* Glucose 6-phosphatase
- D. Phosphoglucomutase
- E. UDPG-glycogen transglucosylase

Recall of Isolated Fact: Which of the following vitamins or minerals is involved in clotting factor synthesis?

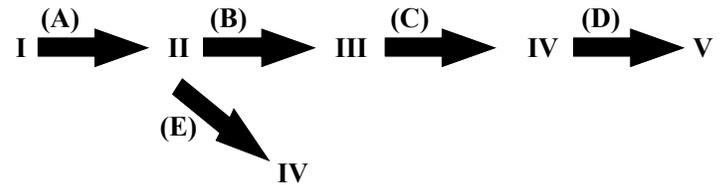
Application of Knowledge: A 70-year-old widower has ecchymoses, perifollicular petechiae, and swelling of the gingiva. His diet consists primarily of cola and hot dogs. Which of the following vitamins or minerals is most likely to be deficient?

- A. Vitamin A
- B. Vitamin B₁
- C. Vitamin B₂
- D. Vitamin B₆
- E.* Vitamin C
- F. Vitamin D
- G. Vitamin E
- H. Vitamin K

- I. Biotin
- J. Copper
- K.* Folate
- L. Iodine
- M. Iron
- N. Magnesium
- O. Niacin
- P. Zinc

Alternate options (beriberi; kwashiorkor; pellagra; rickets; scurvy) seem less appropriate.

In the branched metabolic pathway, a different single enzyme catalyzes each of the individual steps. The enzyme that would be expected to be most severely inhibited by compound V is enzyme



- A. A
- B.* B
- C. C
- D. D
- E. E

A 25-year-old woman is pregnant at 8 weeks' gestation. Her oral intake of calcium is inadequate. If she does not receive supplements which of the following will be the primary source of calcium for the developing fetus

- A.* cancellous bone
- B. liver
- C. parathyroid glands
- D. renal tubules
- E. small intestine

Genetics

A male neonate has severe hemolysis and circulatory failure. The 26-year-old mother is Rh negative and had two previous second-trimester abortions. This condition could have been prevented by administration of which of the following to the mother?

- A. Anti-RhD IgG during the most recent pregnancy
- B.* Anti-RhD IgG on termination of each of the first two pregnancies
- C. Anti-RhD IgM during the most recent pregnancy
- D. Anti-RhD IgM on termination of the first pregnancy

Genes on the bacterial chromosome have the following linkages in conjugal transfer: x and y, 25% of the time; y and z, 50% of the time. If the gene order is x-y-z, approximately what percentage of the time will x and z be transferred together?

- A. 1%
- B. 5%
- C. 13%
- D. 20%
- E.* 40%

Hereditary hyperammonemia, characterized by a grossly abnormal EEG and an increased blood ammonium concentration, is most likely to be caused by a deficiency of which of the following enzymes?

- A. Asparagine synthetase
- B.* Carbamoyl phosphate synthetase I
- C. Fumarase
- D. Glutamate-oxaloacetate aminotransferase
- E. Glutaminase

Histology/Cell Biology

Insert description of "action". Which of the following cellular organelles is most directly involved?

- A. Golgi complex
- B. Lysosome
- C. Peroxisome
- D. Mitochondria
- E. Nuclear envelope
- F. Secretory (zymogenic) granule
- G. Rough endoplasmic reticulum
- H. Smooth endoplasmic reticulum

Lab Vignette

Several contiguous cells are labeled with a fluorescent dye that cannot cross cell membranes. One cell is experimentally bleached with a light that destroys the dye but soon recovers dye fluorescence. This recovery is best explained by the presence of which of the following structures between the bleached cell and its fluorescent neighbors?

- A. Basal lamina
- B. Desmosomes (maculae adherentes)
- C.* Gap junctions
- D. Glycosaminoglycans
- E. Tight junctions (zonulae occludentes)

Microbiology

Template for Microbiology Items:

Item Stem (Patient Vignette): (Describe a patient with a problem)

Lead-In: Infection with which of the following organisms is most likely?

Options: (List of pathogens)

At a banquet, the menu included fried chicken, home fried potatoes, peas, chocolate eclairs, and coffee. Within 2 hours, most of the diners became violently ill, with nausea, vomiting, and abdominal pain. Analysis of the contaminated food is most likely to yield large numbers of which of the following organisms?

- A. *Escherichia coli*
- B. *Proteus mirabilis*
- C. *Salmonella typhimurium*
- D.* *Staphylococcus aureus*
- E. *Streptococcus faecalis*
- F. *Enterococcus*

Insert patient vignette describing a person with a problem. Which of the following toxins is most likely to be involved in pathogenesis?

- A. Botulinum toxin
- B. Diphtheria toxin
- C. Pertussis toxin
- D. Shiga toxin
- E. Tetanus toxin
- F. Cholera enterotoxin
- G. *Clostridium difficile* enterotoxin
- H. *Escherichia coli* heat-stable enterotoxin
- I. *Staphylococcus aureus* enterotoxin

Insert patient vignette describing a patient with a problem. Which of the following is the most appropriate therapeutic agent?

- A. Acyclovir
- B. Amphotericin B
- C. Erythromycin
- D. Ganciclovir
- E. Gentamicin
- F. Ketoconazole
- G. Miconazole
- H. Pyrantel pamoate
- I. Pyrazinamide
- J. Rifampin
- K. Trimethoprim-sulfamethoxazole
- L. Vancomycin
- M. Zidovudine (AZT)

Isolated Fact Item: Which of the following pathogens is an encapsulated gram-positive organism that usually grows in pairs or short chains? (Same option list as below) Answer: T

Application of Knowledge Item: A 7-year-old girl has a high fever and a sore throat. There is pharyngeal redness, a swollen right tonsil with a creamy exudate, and painful right submandibular lymphadenopathy. Throat culture on blood agar yields numerous small α -hemolytic colonies that are inhibited by bacitracin. Which of the following pathogens is most likely to have caused her illness? Answer: U

- A. Adenovirus
- B. *Aspergillus fumigatus*
- C. *Bacillus anthracis*
- D. *Candida albicans*
- E. *Chlamydia psittaci*
- F. *Coccidioides immitis*
- G. Coronavirus
- H. *Corynebacterium diphtheriae*
- I. *Coxiella burnetii*
- J. Cocksackievirus
- K. Epstein-Barr virus
- L. *Haemophilus influenzae*
- M. *Histoplasma capsulatum*
- N. *Mycobacterium tuberculosis*
- O. *Mycoplasma pneumoniae*
- P. *Neisseria gonorrhoeae*
- Q. *Neisseria meningitidis*
- R. *Pneumocystis carinii*
- S. Rhinovirus
- T. *Streptococcus pneumoniae*
- U. *Streptococcus pyogenes* (group A)

Sample Item Set

A 2-year-old boy has had recurrent infections since he was 6 months old. Serum complement concentrations, phagocytic function, and bactericidal of neutrophils are normal. A skin test with *Candida* antigens results in 2 cm of induration by 48 hours. Which of the following provides the best explanation for the time of onset of the disease at 6 months of age?

- A. Development of a viral infection
- B. Exposure to rare bacteria
- C. Loss of passive immunity from the mother
- D. A maturation defect in the thymus
- E. Transplacental transfer of IgM antibodies

Which of the following laboratory tests is most likely to be abnormal?

- A. Activity of myeloperoxidase
- B. CD4/CD8 T lymphocyte ratio
- C. Sweat chloride concentration
- D. Fc receptors on macrophages
- E. Serum IgG concentration

Neuroscience (Neuroanatomy and Neuropathology)

Template for Neuroanatomy Items

Item Stem (Patient Vignette): (Describe a patient with a problem)

Lead-In: An abnormality is most likely to be present at which of the following locations?

Options: (List of sites)

Insert patient vignette describing a patient with a problem. Which of the following cranial nerves is the most likely site of the underlying lesion?

- | | |
|---------------------|----------------------------|
| A. Olfactory nerve | G. Facial nerve |
| B. Optic nerve | H. Vestibulocochlear nerve |
| C. Oculomotor nerve | I. Glossopharyngeal nerve |
| D. Trochlear nerve | J. Vagus nerve |
| E. Trigeminal nerve | K. Spinal accessory nerve |
| F. Abducens nerve | L. Hypoglossal nerve |

Insert patient vignette describing a patient with a neurologic abnormality. Which of the following branches of the brachial plexus is most likely to be affected?

- | | |
|----------------------|---------------------|
| A. Axillary | G. Musculocutaneous |
| B. Dorsal scapular | H. Radial |
| C. Suprascapular | I. Long thoracic |
| D. Upper subscapular | J. Thoracodorsal |
| E. Lower subscapular | K. Ulnar |
| F. Median | |

Insert patient vignette describing a patient with a problem. Which of the following components of the motor system is most likely to be involved?

- A. Basal ganglia
- B. Cerebellar hemisphere
- C. Cerebellar vermis
- D. Motor relay of the thalamus
- E. Premotor area
- F. Sensory motor cortex
- G. Superior colliculus/pretectal area
- H. Supplementary motor area
- I. Ventral horn

Insert patient vignette describing a patient with a problem. The most likely cause is occlusion of which of the following arteries?

- A. Left anterior cerebral
- B. Right anterior cerebral
- C. Left middle cerebral
- D. Right middle cerebral
- E. Left posterior cerebral
- F. Right posterior cerebral
- G. Left lenticulostriate
- H. Right lenticulostriate

Insert patient vignette describing a patient with a problem. Which of the following is the most likely diagnosis?

- A. Amyotrophic lateral sclerosis
- B. Huntington's disease
- C. Multiple sclerosis
- D. Parkinson's disease
- E. Poliomyelitis
- F. Polyneuropathy
- G. Retinitis pigmentosa

Sample Item Set

An unresponsive 58-year-old woman is brought to the emergency department after collapsing at a local shopping mall. Her family reports that she felt well that morning but developed a progressively severe headache. She has had hypertension and atrial fibrillation and is taking an antihypertensive medication and an oral anticoagulant. Her blood pressure is 220/130 mm Hg and she has apnea alternating with hyperpnea. She responds only to noxious stimuli with extensor posturing involving the right arm and leg. Funduscopic examination shows papilledema involving the left optic disc. Pupils are 3.0/7.0 (R/L) with no reaction to light on the left. There is a left gaze preference. There is diffuse hyperreflexia, R > L, and bilateral Babinski signs are present.

1. The dilated, unreactive left pupil is most consistent with injury to which of the following structures on the left?
A. Optic nerve
B. Optic tract
C.* Oculomotor nerve
D. Lateral geniculate nucleus
E. Superior colliculus
2. The extensor posturing on the right is most consistent with injury to which of the following areas of the brain on the left?
A. Telencephalon
B. Diencephalon
C.* Midbrain
D. Pons
E. Medulla
3. Which of the following best describes her respiratory pattern?
A.* Cheyne-Stokes
B. Central neurogenic hyperventilation
C. Apneustic
D. Ataxic
4. Which of the following herniation syndromes is most consistent with her clinical presentation?
A. Cingulate gyrus beneath the falx
B.* Temporal lobe uncus across the tentorium
C. Diencephalon through the tentorial notch
D. Brain stem through the tentorial notch
E. Cerebellar tonsils through the foramen magnum

Pathology

Insert description of "action". Which of the following cellular organelles are most directly involved?

- A. Golgi complex
- B. Lysosomes
- C. Peroxisomes
- D. Mitochondria
- E. Nuclear envelope
- F. Secretory (zymogenic) granules
- G. Rough endoplasmic reticulum
- H. Smooth endoplasmic reticulum

Insert patient vignette describing a patient with a problem. Which of the following is the most likely finding on examination of tissue obtained on renal biopsy?

- A. Acute necrotizing vasculitis
- B. Amyloid deposits
- C. Cortical necrosis
- D. Glomerular mesangial IgA
- E. Glomerular peripheral IgG
- F. Granulomas
- G. Interstitial neutrophils
- H. Nodular glomerulosclerosis
- I. Regenerating tubules
- J. Tubular uric acid crystals

Insert patient vignette describing a patient with a problem. Which of the following endocrine abnormalities is most likely to be present?

- A. Adrenal cortical adenoma
- B. Adrenal cortical carcinoma
- C. Adrenal hyperplasia
- D. Adrenal necrosis
- E. Adrenal neuroblastoma
- F. Adrenal pheochromocytoma
- G. Pancreatic islet cell adenoma
- H. Parathyroid adenoma
- I. Parathyroid carcinoma
- J. Parathyroid hyperplasia

A 32-year-old man dies 4 days after fracturing his femur in an automobile collision. Examination of his brain at autopsy shows widespread petechiae in the cerebral white matter. Which additional autopsy finding is most likely?

- A. Adult respiratory distress syndrome
- B. Contrecoup injury
- C.* Fat embolization
- D. Septicemia
- E. Subdural hematoma

Autopsy of a 24-year-old woman shows pleuritis, membranous thickening of glomerular capillary walls, concentric rings of collagen around splenic arterioles, and excrescences on the underside of the mitral valve. Which of the following findings is most likely on evaluation of blood from this woman?

- A.* Antinuclear antibody
- B. Increased C3 concentration
- C. Lymphocytosis
- D. Monoclonal gammopathy
- E. Positive bacterial culture

A patient with hepatitis B dies 9 days after the onset of symptoms. Which of the following findings is most likely on microscopic examination of his liver?

- A. Diffuse fatty vacuolization with minimal necrosis
- B. Diffuse fibrosis with nodularity
- C. Limited peripheral zonal necrosis
- D.* Widespread hepatocellular necrosis
- E. Widespread neutrophilic infiltration in lobules

Pathology Mechanisms

Template for Mechanisms Items

Item Stem (Patient Vignette): (Describe a patient with a problem)

Lead-In: Which of the following mechanisms is the most likely cause of the patient's findings?

Options: (List of mechanisms)

Insert patient vignette which describes a patient with a problem. Which of the following is the most likely cause of the myocardial disease?

- A. Alcohol toxicity
- B. Cardiac amyloidosis
- C. Endomyocardial fibrosis
- D. Hemochromatosis
- E. Hypertrophic cardiomyopathy
- F. Löffler's endocarditis
- G. Postviral myocarditis
- H. Sarcoidosis
- I. South American trypanosomiasis (Chagas' disease)
- J. Vitamin B₁ (thiamine) deficiency

Insert patient vignette describing a patient with jaundice. Which of the following is the most likely cause of the jaundice?

- A. Alcoholic hepatic disease
- B. Drug reaction
- C. Dubin-Johnson syndrome
- D. Gilbert's syndrome
- E. Hemolytic jaundice
- F. Intrahepatic ductal atresia
- G. Obstructive jaundice
- H. Primary biliary cirrhosis
- I. Viral hepatitis

Insert patient vignette describing a patient with a hemostatic abnormality. Which of the following is the most likely cause of the hemostatic abnormality?

- A. Acute disseminated intravascular coagulation
- B. Factor V (proaccelerin) deficiency
- C. Factor VII (proconvertin) deficiency
- D. Hemophilia A
- E. Hemophilia B
- F. Idiopathic thrombocytopenia purpura
- G. Pulmonary thromboembolism
- H. von Willebrand's disease

Insert patient vignette describing a patient with a problem. Which of the following is the most likely gastrointestinal lesion?

- A. Candidal esophagitis
- B. Diverticulitis
- C. Hiatal hernia
- D. Peptic ulcer
- E. Pseudomembranous enterocolitis
- F. Pyloric stenosis
- G. Regional enteritis
- H. Subphrenic abscess
- I. Ulcerative colitis

Pathology Diagnosis

Template for Diagnosis Items

Item Stem (Patient Vignette): (Describe a patient with a problem)

Lead-In: Which of the following is the most likely diagnosis?

Options: (List of diagnoses)

Insert patient vignette to describe a patient with a histologic abnormality. Which of the following is the most likely thyroid disorder?

- A. Chronic autoimmune (Hashimoto's) thyroiditis
- B. Graves' disease
- C. Myxedema after ^{131}I therapy
- D. Subacute thyroiditis
- E. Well-differentiated thyroid carcinoma

Insert patient vignette describing a patient with a problem. Which of the following is the most likely diagnosis?

- A. Acute lymphoblastic leukemia
- B. Acute myelogenous leukemia
- C. AIDS
- D. Chronic lymphocytic leukemia
- E. Ewing's sarcoma
- F. Follicular lymphoma
- G. Hodgkin's disease
- H. Idiopathic thrombocytopenic purpura
- I. Multiple myeloma
- J. Polycythemia vera
- K. Histiocytosis X
- L. T-cell lymphoma

Insert patient vignette describing a patient with heart disease. Which of the following is the most likely diagnosis?

- A. Acute viral myocarditis
- B. Chagas' disease
- C. Dilated cardiomyopathy
- D. Giant cell myocarditis
- E. Hemochromatosis
- F. Hypertensive cardiomyopathy
- G. Hypertrophic cardiomyopathy
- H. Ischemic cardiomyopathy
- I. Pompe's disease
- J. Restrictive cardiomyopathy

Insert patient vignette to describe a patient with a problem. Which of the following cell types is the most likely to be involved in the lesion?

- A. Basophil
- B. Endothelial cell
- C. Eosinophil
- D. Lymphocyte
- E. Mesothelial cell
- F. Monocyte
- G. Neutrophil
- H. Plasma cell

Insert patient vignette describing a patient with renal disease. Which of the following is the most likely diagnosis?

- A. Acute poststreptococcal glomerulonephritis
- B. Focal segmental sclerosis
- C. Goodpasture's syndrome
- D. IgA nephropathy
- E. Lupus nephritis
- F. Membranoproliferative glomerulonephritis
- G. Membranous glomerulonephritis
- H. Minimal change nephropathy

A 21-year-old man has weight loss and severe intermittent bloody diarrhea. Barium enema and colonoscopy show multiple ulcers and inflammatory changes extending from the rectum to the mid-transverse colon. Biopsy specimens taken from multiple sites show acute and chronic inflammation restricted to the mucosa. Which of the following is the most likely diagnosis?

- A. AIDS-associated gastroenteritis
- B. Amebiasis
- C.* Crohn's disease
- D. *Clostridium difficile*-associated colitis
- E. *Escherichia coli*-associated colitis
- F. Ischemic colitis
- G. *Salmonella* gastroenteritis
- H. Ulcerative colitis

A 38-year-old woman has congestive heart failure, premature ventricular contractions, and repeated episodes of ventricular tachycardia. Her blood pressure is normal. Her heart is markedly enlarged. There are no murmurs; coronary angiography is normal. Which of the following is the most likely diagnosis?

- A. Acute rheumatic fever
- B. Congenital fibroelastosis
- C. Constrictive pericarditis
- D.* Myocardial infarction
- E. Primary cardiomyopathy

A 74-year-old man has colicky abdominal pain in the left lower quadrant, leukocytosis, and fever. He has not had diarrhea or constipation. Which of the following is the most likely diagnosis?

- A. Carcinoma of the sigmoid colon
- B.* Diverticulitis
- C. Familial adenomatous polyposis
- D. Ulcerative colitis
- E. Villous adenoma of the upper rectum

Pathophysiology

Template for Pathophysiology Items - Additional Findings

Item Stem (Patient Vignette): (Describe a patient with a problem)

Lead-In: Which of the following additional findings is most likely to be present?

Options: (List of findings)

Laboratory tests on an edematous 35-year-old man show a normal serum concentration of complement and an increased serum concentration of cholesterol. Urinalysis show proteinuria (4+), 0-5 erythrocytes/hpf, and several hyaline casts. Which of the following findings is most likely on renal biopsy?

- A. Acute poststreptococcal (proliferative) glomerulonephritis
- B. Membranoproliferative glomerulonephritis
- C.* Membranous glomerulonephritis
- D. Minimal change disease
- E. Rapidly progressive glomerulonephritis

A 16-year-old boy is undergoing evaluation for jaundice. Laboratory studies show normal hepatic enzyme activities, a negative direct antiglobulin test, increased mean corpuscular hemoglobin concentration, and increased osmotic fragility of erythrocytes. Which of the following types of erythrocyte is most likely to be seen on a peripheral blood smear?

- A. Ovalocyte
- B. Schistocyte
- C.* Spherocyte
- D. Target cell
- E. Teardrop cell

An asymptomatic 50-year-old woman has hypertension. Urinary excretion of catecholamines is increased. A CT scan shows a suprarenal mass. Which of the following findings is most likely on microscopic examination of the resected mass?

- A. Benign neoplasm of the adrenal cortex
- B.* Benign neoplasm of the adrenal medulla
- C. Malignant neoplasm of the adrenal cortex
- D. Malignant neoplasm of the adrenal medulla
- E. Diffuse hyperplasia of the adrenal cortex
- F. Diffuse hypoplasia of the adrenal medulla

Insert patient vignette describing a patient with a problem. Which of the following is the most likely pulmonary finding at autopsy?

- A. Soft with fluffy consistency, multiple blebs over the surface
- B. Very heavy; frothy and bloody fluid exudes freely from the cut surface
- C. Diffuse fibrotic nodules; whorled appearance replacing normal lung parenchyma
- D. Liver-like consistency in left lower lobe; microscopically contains fibrin and neutrophils in the alveoli
- E. Normal-appearing; large Y-shaped, coiled, laminated clot almost filling both pulmonary arteries

Insert patient vignette describing a patient with a problem. Which of the following sets of findings is most likely?

<u>Effective circulating volume</u>	<u>Extracellular fluid volume</u>	<u>Plasma volume</u>	<u>Urine Na⁺ excretion</u>
A. Decreased	decreased	decreased	decreased
B. Decreased	increased	decreased	decreased
C. Decreased	increased	increased	decreased
D. Increased	increased	increased	increased

Insert patient vignette describing a patient with a problem. Which of the following digestive enzymes or cofactors is most likely to be involved?

- A. Amylase
- B. Chymotrypsin
- C. Colipase
- D. Enterokinase
- E. Lactase
- F. Lipase
- G. Pepsin
- H. Sucrase
- I. Trypsin

Insert patient vignette describing a patient with a problem. Which of the following laboratory findings is the most likely?

- | | |
|--------------------|------------------|
| A. Hypercalcemia | E. Hypernatremia |
| B. Hypocalcemia | F. Hyponatremia |
| C. Hypermagnesemia | G. Hyperkalemia |
| D. Hypomagnesemia | H. Hypokalemia |

Insert patient vignette describing a patient with a problem. Which of the following is the most likely thyroid function profile?

	<u>Thyroxine (T₄)</u>	<u>Triiodothyronine (T₃)</u>	<u>Resin Uptake</u>	<u>Thyroid-stimulating Hormone</u>	<u>T₃</u>
A.	↓	↓		↓	↓
B.	↓	normal		↑	↓
C.	↓	normal		↑	normal
D.	↓	↑		normal	normal
E.	↑	↑		↓	↑
F.	↑	normal		↑	↑
G.	normal	normal		normal	normal

Insert patient vignette describing a patient with electrocardiographic findings. Which of the following cardiac arrhythmias is most likely to be present?

- | | |
|---|---------------------------------------|
| A. Atrial fibrillation | F. Premature atrial contractions |
| B. Atrial paroxysmal tachycardia | G. Premature ventricular contractions |
| C. First-degree incomplete heart block | H. Sinus arrhythmia |
| D. Second-degree incomplete heart block | I. Ventricular fibrillation |
| E. Third-degree complete heart block | J. Ventricular paroxysmal tachycardia |

Insert patient vignette describing a patient with cardiac abnormalities (A 4-week-old infant has a loud systolic murmur with a systolic thrill; otherwise, the infant appears healthy. There is no cyanosis). Which of following congenital heart diseases is most likely to be present?

- A. Atrial septal defect
- B. Coarctation of the aorta
- C. Complete transposition of the great arteries
- D. Endocardial cushion defect
- E. Patent ductus arteriosus
- F. Pulmonic valve stenosis
- G. Tetralogy of Fallot
- H. Ventricular septal defect

Pharmacology

Template for Adverse Effects Pharmacology – "Guess My Drug"

Item Stem (Patient Vignette): (Describe a patient with adverse drug effect)

Lead-In: Which of the following drugs has the patient most likely been taking?

Options: (List of drugs)

A patient in the emergency department cannot remember which "heart drug" he is taking. He says he has "ringing in his ears." His heart rate is greater than 80/min. The ECG shows prolonged PR and QRS intervals. Which of the following drugs has the patient most likely been taking?

- A. Digoxin
- B. Lidocaine
- C. Phenytoin
- D. Propranolol
- E.* Quinidine

Insert patient vignette describing a patient with an adverse drug effect. Which of the following is the most likely to have caused the adverse effect?

- | | |
|-------------------------|-------------------|
| A. Acetaminophen | J. Nalidixic acid |
| B. Amiodarone | K. Nitrofurantoin |
| C. ACE inhibitors | L. Penicillin |
| D. Aspirin | M. Prednisone |
| E. Atenolol | N. Procainamide |
| F. Bleomycin | O. Propranolol |
| G. Cytosine arabinoside | P. Sulfasalazine |
| H. Furosemide | Q. Tetracycline |
| I. Metronidazole | R. Verapamil |

Drug Y has a volume of distribution (V_d) of 75 L in both younger and older adult men. In younger adults, it has a clearance rate of 15L/h, 50% of which is via the liver and 50% via the kidneys. For younger men, the maintenance regimen is 100 mg every 6 hours. Which of the following regimens will produce essentially the same steady-state concentration in an older man, whose creatinine clearance is half that of younger men, but whose hepatic function is unimpaired?

- A. 75 mg every 3 hours
- B.* 75 mg every 6 hours
- C. 75 mg every 9 hours
- D. 100 mg every 3 hours
- E. 100 mg every 6 hours
- F. 100 mg every 12 hours

A 24-year-old sexually active woman reports increased vaginal discharge. Pelvic examination shows a green frothy discharge. Microscopic examination of a wet mount of the discharge shows motile unicellular organisms 10-30 μ in length. Which of the following features of this disorder best explains why the infection can be effectively treated with metronidazole?

- A. Facultatively aerobic
- B. Microaerophilic
- C. Strictly aerobic
- D.* Strictly anaerobic

Template for Site/Mechanism of Action Items – Pharmacology

Item Stem (Patient Vignette): (Describe a patient who requires pharmacotherapy)

Lead-In: A drug with which of the following sites/mechanisms of action is most likely to be effective?

Options: (List of mechanisms of action) OR (List of sites of action)

Pharmacotherapy

Template for Pharmacotherapy Items

Item Stem (Patient Vignette): (Describe a patient who requires pharmacotherapy)

Lead-In: Administration of which of the following is most appropriate?

Options: (List of drugs)

A 40-year-old black man has the sudden onset of severe headache, dizziness, and vomiting. His blood pressure is 260/130 mm Hg; he has encephalopathy and grade IV retinopathy. Administration of which of the following is most appropriate?

- | | |
|-----------------------|------------------------------|
| A. Alpha agonists | G. Central sympatholytics |
| B. Alpha blockers | H. Direct vasodilators |
| C. ACE inhibitors | I. Negative inotropic agents |
| D. Beta agonists | J. Thiazide diuretics |
| E. Beta blockers | K. Vasoconstrictors |
| F. Cardiac glycosides | |

Insert patient vignette describing a patient who requires pharmacotherapy. Which of the following is the most appropriate initial drug therapy?

- | | |
|----------------|------------------|
| A. Adenosine | G. Naloxone |
| B. Aspirin | H. Prednisone |
| C. Caffeine | I. Propranolol |
| D. Epinephrine | J. Quinidine |
| E. Insulin | K. Streptokinase |
| F. Lidocaine | |

Insert patient vignette describing a patient who requires pharmacotherapy. Which of the following is the most appropriate therapeutic agent?

- A. Carbamazepine
- B. Dextroamphetamine
- C. Ethosuximide
- D. Haloperidol
- E. L-dopa/carbidopa
- F. Lithium carbonate
- G. Phenobarbital
- H. Primidone
- I. Propranolol
- J. Pyridostigmine

Insert patient vignette describing a patient who requires pharmacotherapy. Which of the following is the most appropriate therapeutic agent?

- A. Atropine
- B. Bismuth subsalicylate
- C. Cimetidine
- D. Diphenoxylate
- E. Kaolin
- F. Milk of magnesia
- G. Misoprostol
- H. Omeprazole
- I. Ranitidine
- J. Sucralfate

Insert patient vignette describing a patient who requires pharmacotherapy. Which of the following is the most appropriate diuretic?

- A. Carbonic anhydrase inhibitor
- B. High-ceiling or loop
- C. Nonsteroidal potassium-sparing
- D. Osmotic
- E. Steroidal potassium-sparing
- F. Thiazide
- G. Xanthine

Physiology

An anesthetized patient is being mechanically ventilated. Initial arterial blood gas values are normal. If the ventilation is decreased, which of the following best describes arterial PCO_2 and pH.

Arterial PCO_2	pH
A. Decrease	decrease
B. Decrease	increase
C.* Decrease	no change
D. Increase	decrease
E. Increase	increase
F. Increase	no change

Insert patient vignette describing a patient with a problem. Which of the following substances is most likely involved?

- | | |
|-------------------------------|------------------------|
| A. ADH (vasopressin) | E. Bradykinin |
| B. Aldosterone | F. Calcitonin |
| C. Angiotensin | G. Parathyroid hormone |
| D. Atrial natriuretic peptide | H. Renin |

A 22-year-old man with a 3-week history of polyuria and polydipsia has had nausea, vomiting, and decreased responsiveness for the past 12 hours. Urinalysis shows 4+ glucose and 4+ ketones. Which of the following sets of arterial blood gas findings is most likely?

	pH	PO_2 (mm Hg)	PCO_2 (mmHg)	HCO_3 (mEq/L)
A.	7.15	98	33	11
B.*	7.15	98	24	8
C.	7.30	56	80	38
D.	7.40	100	40	25
E.	7.50	100	33	25
F.	7.50	100	24	18
G.	7.50	56	33	25

Insert patient vignette describing a patient with a problem. Which of the following humoral substances is most likely involved?

- A. Acetylcholine
- B. Adenosine
- C. Aldosterone
- D. Bradykinin
- E. Epinephrine
- F. Norepinephrine
- G. Prostacyclin (PGI₂)
- H. Prostaglandin E₂
- I. Prostaglandin F₂"
- J. Serotonin
- K. Thromboxane A₂

Integrative Item Sets

Template for Integrative Indications Items: Microbiology/Pharmacology

Item Stem (Patient Vignette) (Describe a patient with a problem)

Lead-In: Infection with which of the following pathogens is most likely?

Options: (List of pathogens)

Lead-In: Administration of which of the following drugs is most appropriate?

Options: (List of drugs)

A 40-year-old man with AIDS has a 1-week history of low-grade fever and lethargy. Temperature is 38 C (100.4 F), pulse is 90/min, and blood pressure is 110/70 mm Hg. There is slight resistance to passive flexion of the neck. Laboratory studies of cerebrospinal fluid show:

Opening pressure	210 mm Hg
Appearance	straw-colored
Leukocyte count	400/mm ³ (100% lymphocytes)
Erythrocyte count	50/mm ³
Glucose	30 mg/dL
Protein	100 mg/dL
India ink preparation	encapsulated yeast

1. Infection with which of the following pathogens is most likely?
 - A. *Blastomyces*
 - B. *Cryptococcus*
 - C. *Histoplasma*
 - D. *Toxoplasma gondii*
 - E. *Treponema pallidum*
2. Administration of which of the following drugs is most appropriate?
 - A. Amphotericin
 - B. Ketoconazole
 - C. Miconazole
 - D. Nystatin
 - E. Trimethoprim-sulfamethoxazole

Sample Item Set

A 34-year-old woman has had severe watery diarrhea for the past 4 days. Two months ago, she had infectious mononucleosis. She abuses drugs intravenously and is seropositive for HIV. Physical examination shows dehydration and marked muscle weakness.

1. Which of the following laboratory abnormalities is most likely?
 - A. Decreased serum K^+ concentration
 - B. Decreased serum Ca^{2+} concentration
 - C. Increased serum HCO_3^- concentration
 - D.* Increased serum Na^+ concentration
 - E. Increased serum pH
2. In evaluating the cause of the diarrhea, which of the following studies is most appropriate?
 - A. Colonic biopsy to identify *Giardia lamblia*
 - B. Culture of the oral mucosa for *Candida albicans*
 - C. Duodenal biopsy to identify *Entamoeba histolytica*
 - D. Gastric aspirate to identify *Mycobacterium avium-intracellulare*
 - E.* Stool specimen to identify *Cryptosporidium*
3. Further studies to evaluate her HIV infection show a helper to suppressor T lymphocyte ratio of 0.3. Which of the following actions of HIV best explains this ratio?
 - A. Induction of helper T lymphocyte proliferation
 - B. Induction of suppressor T lymphocyte proliferation
 - C.* Infection of cells with CD4 receptor
 - D. Infection of macrophages
 - E. Stimulation of leukotriene synthesis

Template for Pathology/Pharmacology Items

Item Stem (Patient Vignette): (Describe a patient with a problem)

Lead-In: Which of the following is the most likely diagnosis?

Options: (List of diagnoses)

Next Lead-In: Administration of which of the following drugs is most appropriate?

Next List of Options: (List of drugs)

Sample Item Set

1. A 62-year-old man with alcohol dependence is admitted to the hospital for transurethral resection of the prostate. The following morning, while being transported to the operating room, he has two generalized seizures within 5 minutes. Neurologic examination shows no focal abnormalities. Which of the following is the most likely diagnosis?
 - A. * Alcohol withdrawal
 - B. Korsakoff's syndrome
 - C. Partial complex seizure
 - D. Wernicke's encephalopathy
2. The most appropriate management is intravenous administration of which of the following drugs?
 - A. Diazepam
 - B. Haloperidol
 - C. Phenobarbital
 - D. Phenytoin
 - E. Valproate

Sample Item Set

A 2-year-old boy has had recurrent infections since he was 6 months old. Serum complement concentrations, phagocytic function, and bactericidal activity of neutrophils are normal. A skin test with *Candida* antigens results in 2 cm of induration by 48 hours. Which of the following provides the best explanation for the time of onset of the disease at 6 months of age?

- A. Development of a viral infection
- B. Exposure to rare bacteria
- C. Loss of passive immunity from the mother
- D. A maturation defect in the thymus
- E. Transplacental transfer of IgM antibodies

Which of the following laboratory tests is most likely to be abnormal?

- A. Activity of myeloperoxidase
- B. CD4/CD8 T lymphocyte ratio
- C. Sweat chloride concentration
- D. Fc receptors on macrophages
- E. Serum IgG concentration

Template for Integrative Physiology/Pharmacology Items

Item Stem (Patient Vignette): (Describe a patient with a problem)

Lead-In: Administration of a drug aimed at achieving which of the following effects is most appropriate? *OR*
Administration of {specific drug} is most likely to have which of the following effects?

Options: (List of physiological effects)

Sample Item-Writing Option Lists For the Clinical Sciences

Components of a Patient Vignette

Age, Gender (eg, A 45-year-old man)

Site of Care (eg, comes to the emergency department)

Presenting Complaint (eg, because of headache)

Duration (eg, that has continued for 2 days).

Patient History (with Family History?)

Physical Findings

+/- Results of Diagnostic Studies

+/- Initial Treatment, Subsequent Findings, Etc.

Abdominal Mass

- | | |
|-------------------------------|-------------------------------|
| A. Abdominal abscess | J. Hernia |
| B. Acute lymphocytic leukemia | K. Hydatid cyst |
| C. Carcinoma of the colon | L. Inflammatory bowel disease |
| D. Carcinoma of the ovary | M. Neuroblastoma |
| E. Choledochal cyst | N. Pancreatic pseudocyst |
| F. Constipation | O. Polycystic kidney |
| G. Desmoid tumor | P. Uterine fibroids |
| H. Distended bladder | Q. Volvulus |
| I. Hepatoma | R. Wilms' tumor |

For each patient with an abdominal mass, select the most likely diagnosis.

Abdominal Pain

- A. Abdominal aneurysm
- B. Appendicitis
- C. Bowel obstruction
- D. Cholecystitis
- E. Colon cancer
- F. Constipation
- G. Diverticulitis
- H. Ectopic pregnancy – ruptured
- I. Endometriosis
- J. Hernia
- K. Kidney stone
- L. Mesenteric adenitis
- M. Mesenteric artery thrombosis
- N. Ovarian cyst – ruptured
- O. Pancreatitis
- P. Pelvic inflammatory disease
- Q. Peptic ulcer disease
- R. Perforated peptic ulcer
- S. Pyelonephritis
- T. Torsion

For each patient with abdominal pain, select the most likely diagnosis.

Altered Mental Status

- A. Brain abscess
- B. Cerebrovascular occlusion/
transient ischemic episode
- C. Dehydration
- D. Drug overdose/drug toxicity
- E. Hepatic encephalopathy
- F. Hypercalcemia
- G. Hyperglycemia
- H. Hyperkalemia
- I. Hypertensive encephalopathy
- J. Hypoglycemia
- K. Hyponatremia
- L. Hypothyroidism
- M. Multi-infarct dementia
- N. Primary degenerative dementia, Alzheimer type
- O. Seizure
- P. Sepsis
- Q. Uremia
- R. Wernicke's encephalopathy

For each patient with altered mental status, select the most likely diagnosis.

Anemia

- A. Acute myeloblastic leukemia
- B. Aplastic anemia
- C. Chronic lymphocytic leukemia
- D. Drug-induced immune hemolytic anemia
- E. Falciparum malaria
- F. Folate deficiency
- G. Glucose 6-phosphate dehydrogenase deficiency
- H. Hereditary spherocytosis
- I. Iron deficiency anemia
- J. Malignancy metastatic to bone marrow
- K. Microangiopathic hemolytic anemia
- L. Multiple myeloma
- M. Myelofibrosis
- N. Sickle cell disease
- O. Thalassemia minor
- P. Vitamin B₁₂ (cyanocobalamin) deficiency

For each patient with anemia, select the most likely diagnosis.

Back Pain

- A. Ankylosing spondylitis
- B. Intervertebral disc infection
- C. Multiple myeloma
- D. Myofascial pain
- E. Osteoporosis
- F. Spinal stenosis
- G. Spondylolysis
- H. Tuberculosis of the spine

For each patient with back pain, select the most likely diagnosis.

Bizarre Behavior

- A. Acidosis
- B. Acute psychosis
- C. Acute subdural hematoma
- D. Alcohol intoxication
- E. Cocaine ingestion/overdose
- F. Delirium tremens
- G. Depression
- H. Hypoglycemia
- I. Hypoxia
- J. LSD ingestion
- K. Malingering
- L. Meningitis
- M. Reye's syndrome
- N. Steroid psychosis
- O. Subarachnoid hemorrhage
- P. Temporal lobe seizure
- Q. Tricyclic antidepressant overdose

For each patient with bizarre behavior, select the most likely diagnosis.

Breast Lump / Complaint

- A. Breast cyst
- B. Fat necrosis of the breast
- C. Fibroadenoma of the breast
- D. Galactocele
- E. Gynecomastia
- F. Inflammatory carcinoma of the breast
- G. Intraductal papilloma
- H. Lipoma of the breast
- I. Mastodynia
- J. Metastatic carcinoma of the breast
- K. Paget's disease of the breast
- L. Puerperal mastitis
- M. Tietze's syndrome

For each patient with a breast-related problem, select the most likely diagnosis.

Chest Pain

- A. Angina pectoris
- B. Compression fracture of the spine
- C. Dissecting aortic aneurysm
- D. Esophageal spasm
- E. Esophagitis
- F. Herpes zoster
- G. Hyperventilation
- H. Myocardial infarction
- I. Pectoral muscle pain
- J. Pericarditis
- K. Pneumonia
- L. Pneumothorax
- M. Pulmonary embolism

For each patient with chest pain, select the most likely diagnosis.

Diarrhea

- A. Amebiasis
- B. Bacterial gastroenteritis
- C. Carcinoid syndrome
- D. Carcinoma of the colon
- E. Chronic pancreatitis
- F. Crohn's disease
- G. Cryptosporidium infection
- H. Diverticulitis
- I. Dumping syndrome
- J. Gastric lymphoma
- K. Ischemic colitis
- L. Laxative abuse
- M. Pseudomembranous colitis
- N. Short gut syndrome
- O. Sprue
- P. Villous adenoma
- Q. Viral gastroenteritis
- R. Zollinger-Ellison syndrome

For each patient with diarrhea, select the most likely diagnosis.

Fatigue

- A. Acute leukemia
- B. Anemia: chronic disease
- C. Congestive heart failure
- D. Depression
- E. E-B virus infection
- F. Folate deficiency
- G. Glucose 6-phosphate dehydrogenase deficiency
- H. Hereditary spherocytosis
- I. Hypothyroidism
- J. Iron deficiency
- K. Lyme disease
- L. Miliary tuberculosis
- M. Vitamin B₁₂ deficiency

For each patient with fatigue, select the most likely diagnosis.

Fever/Adults

- A. Appendicitis
- B. Aspirin intoxication
- C. Cellulitis
- D. Crohn's disease
- E. Gastritis
- F. Hodgkin's disease
- G. Infectious mononucleosis
- H. Meningitis
- I. Narcotic overdose
- J. Pancreatitis
- K. Pneumonia
- L. Prostatitis
- M. Pulmonary embolism
- N. Pyelonephritis
- O. Sinusitis
- P. Tuberculosis
- Q. Urinary tract infection
- R. Viral respiratory infection

For each patient with fever, select the most likely diagnosis.

Fever in Children

- A. Acute lymphocytic leukemia
- B. Bacterial meningitis
- C. Chickenpox
- D. Drug fever
- E. Gastroenteritis
- F. Juvenile rheumatoid arthritis
- G. Kawasaki disease
- H. Mastoiditis
- I. Measles
- J. Osteomyelitis
- K. Otitis media
- L. Pneumococcal pneumonia
- M. Pyelonephritis
- N. Rheumatic fever
- O. Roseola
- P. Rubella
- Q. Sinusitis
- R. Streptococcal infection

For each patient with fever, select the most likely diagnosis.

Gastrointestinal Bleeding

- A. Amebiasis
- B. Angiodysplasia of the colon
- C. Clostridium difficile colitis
- D. Carcinoma of the colon
- E. Carcinoma of the esophagus
- F. Carcinoma of the stomach
- G. Coagulopathy
- H. Diverticulitis
- I. Epistaxis
- J. Esophageal varices
- K. Hemorrhoids
- L. Inflammatory bowel disease
- M. Ischemic colitis
- N. Mallory-Weiss tear
- O. Peptic ulcer disease
- P. Reflux esophagitis
- Q. Salmonella infection
- R. Shigella infection

For each patient with gastrointestinal bleeding, select the most likely diagnosis.

Headache

- A. Aseptic meningitis
- B. Bacterial meningitis
- C. Brain abscess
- D. Brain metastases
- E. Cerebrovascular occlusion
- F. Cluster headache
- G. Drug-induced headache
- H. Dysfunction of temporomandibular joint
- I. Hypertensive crisis
- J. Migraine
- K. Primary brain tumor
- L. Pseudotumor cerebri
- M. Subacute bacterial endocarditis
- N. Subarachnoid hemorrhage
- O. Temporal arteritis
- P. Tension headache
- Q. Trigeminal neuralgia

For each patient with headache, select the most likely diagnosis.

Leg Pain

- A. Acute arterial insufficiency
- B. Buerger's disease
- C. Causalgia
- D. Cellulitis
- E. Chronic arterial insufficiency
- F. Chronic venous stasis
- G. Dermatomyositis
- H. Diabetic neuropathy
- I. Frostbite
- J. Leriche's syndrome
- K. Lymphedema
- L. Obstruction of the inferior vena cava
- M. Osteomyelitis
- N. Pressure ulcer
- O. Raynaud's disease
- P. Thrombophlebitis
- Q. Varicose veins
- R. Venous insufficiency

For each patient with leg pain, select the most likely diagnosis.

Limb Weakness

- A. Brachial plexopathy
- B. C1-2 subluxation
- C. C7-T1 subluxation
- D. Encephalopathy
- E. Epidural abscess
- F. Fracture of the cervical spine
- G. Guillain-Barré syndrome
- H. Lead neuropathy
- I. Migraine
- J. Multiple sclerosis
- K. Myasthenia gravis
- L. Osteoarthritis
- M. Poliomyelitis
- N. Postictal state
- O. Ruptured cervical disc ischemia
- P. Stroke/transient cerebrovascular
- Q. Subdural hematoma
- R. Vertebral metastases

For each patient with weakness of a limb, select the most likely diagnosis.

Nausea/Vomiting

- A. Bulimia nervosa
- B. Digoxin toxicity
- C. Gastric outlet obstruction
- D. Gastroenteritis
- E. Gastroparesis
- F. Hepatitis
- G. Inflammatory bowel disease
- H. Intracranial hemorrhage
- I. Labyrinthitis
- J. Meningitis
- K. Migraine
- L. Myocardial infarction
- M. Pancreatic carcinoma
- N. Pancreatitis
- O. Pregnancy
- P. Sigmoid volvulus
- Q. Small bowel obstruction
- R. Vasovagal attack

For each patient with nausea and vomiting, select the most likely diagnosis.

Respiratory Problems

- A. Allergic reaction
- B. Aspiration pneumonia
- C. Bronchiectasis
- D. Carcinoma of the lung
- E. Chronic obstructive pulmonary disease
- F. Congestive heart failure
- G. Cystic fibrosis
- H. Drug side effects
- I. Exercise-induced asthma
- J. Foreign body
- K. Gastroesophageal reflux
- L. Mitral stenosis
- M. Occupational asthma
- N. Pleurisy
- O. Pulmonary embolism
- P. Respiratory syncytial virus
- Q. Sarcoidosis
- R. Tuberculosis

For each patient with respiratory symptoms, select the most likely diagnosis.

Shock

- A. Addisonian crisis
- B. Anaphylaxis
- C. Atrial fibrillation
- D. Atrioventricular block
- E. Bleeding peptic ulcer
- F. Carcinoid syndrome
- G. Cardiac tamponade
- H. Cardiomyopathy
- I. Dehydration
- J. Fracture of the pelvis
- K. Myocardial infarction
- L. Pulmonary embolus
- M. Ruptured abdominal aortic aneurysm
- N. Ruptured abscess of the appendix
- O. Ruptured ectopic pregnancy
- P. Ruptured ovarian cyst
- Q. Small bowel obstruction

For each patient with shock, select the most likely diagnosis.

Shortness of Breath

- A. Abscess of the lung
- B. Adult respiratory distress syndrome
- C. Anemia
- D. Asthma
- E. Chronic obstructive pulmonary disease
- F. Congestive heart failure
- G. Diaphragmatic rupture
- H. Foreign body
- I. Hyperventilation
- J. Laryngeal spasm
- K. Mesothelioma
- L. Myasthenia gravis
- M. Pancoast's tumor
- N. Pleural effusion
- O. Pneumothorax
- P. Primary pulmonary hypertension
- Q. Pulmonary embolism
- R. Pulmonary fibrosis

For each patient with shortness of breath, select the most likely diagnosis.

Urinary Problems

- A. Acute cystitis
- B. Acute urinary retention
- C. Carcinoma of the bladder
- D. Carcinoma of the prostate
- E. Diabetes insipidus
- F. Diabetes mellitus
- G. Hypercalcemia
- H. Interstitial cystitis
- I. Neurogenic bladder
- J. Psychogenic polydipsia
- K. Prostatic hyperplasia
- L. Prostatitis
- M. Pyelonephritis
- N. Renal cell carcinoma
- O. Urethral stricture

For each patient with a urinary problem, select the most likely diagnosis.

Urinary Symptoms

- A. Acute urinary retention
- B. Bladder fistula
- C. Carcinoma of the bladder
- D. Cystitis
- E. Drug effect
- F. Endometriosis
- G. Gonorrhea
- H. March hemoglobinuria
- I. Menstruation
- J. Myoglobinuria
- K. Nephrolithiasis
- L. Pneumaturia
- M. Prostatism
- N. Renal carcinoma
- O. Stress incontinence
- P. Syphilis
- Q. Trauma

For each patient with urinary symptoms, select the most likely diagnosis.

Vaginal Bleeding/Pain

- A. Abruptio placentae
- B. Bacterial vaginitis
- C. Breakthrough bleeding
- D. Candidal vaginitis
- E. Cervical carcinoma
- F. Condyloma acuminatum
- G. Ectopic pregnancy
- H. Endometrial cancer
- I. Foreign body
- J. Gonococcal cervicitis
- K. Molar pregnancy
- L. Normal menses
- M. Placenta previa
- N. Threatened abortion
- O. Trichomoniasis
- P. Vesicovaginal fistula
- Q. Vulvar carcinoma

For each patient with vaginal bleeding or pain, select the most likely diagnosis.

Vaginal Discharge/Pruritus

- A. Allergic vaginitis
- B. Atrophic vaginitis
- C. Bacterial vaginitis
- D. Breakthrough bleeding
- E. Candidal vaginitis
- F. Cervical carcinoma
- G. Condyloma acuminatum
- H. Endometrial cancer
- I. Foreign body
- J. Gonococcal cervicitis
- K. Normal menses
- L. Placenta previa
- M. Threatened abortion
- N. Trichomoniasis
- O. Vesicovaginal fistula
- P. Vulvar carcinoma

For each patient with vaginal discharge or pruritus, select the most likely diagnosis.

Wheezing

- A. Angioedema
- B. Asthma
- C. Bronchiectasis
- D. Carcinoid syndrome
- E. Chronic obstructive pulmonary disease
- F. Congestive heart failure
- G. Cystic fibrosis
- H. Endobronchial polyp
- I. Epiglottitis
- J. Esophageal reflux
- K. Foreign body
- L. Goiter
- M. Laryngeal dyskinesia
- N. Pneumonia
- O. Pulmonary edema
- P. Pulmonary embolism
- Q. Respiratory syncytial virus infection
- R. Viral croup

For each patient with wheezing, select the most likely diagnosis.