Item-writing skills:
a) Are essential to compiling valid assessment tools
b) Can prevent your questions from giving away the answers
c) Are Best learned by example
d) All of the above

TEST-ITEM DEVELOPMENT *for* RADIOLOGIC TECHNOLOGY

You will learn:
a) Item format options and development guidelines
b) The 5-step process for turning ideas into items
c) How to assess critical thinking skills
d) All of the above

This book is:
a) A workshop guide and manual
b) A 'how-to' for creating multiple-choice questions
c) An ARRT-copyrighted publication
d) All of the above

Promoting high standards of patient care by recognizing qualified individuals in diagnostic medical imaging, interventional procedures and radiation therapy

The American Registry of Radiologic Technologists®

TEST-ITEM DEVELOPMENT *for* RADIOLOGIC TECHNOLOGY

Authors

Mark Raymond
ARRT Director of Psychometric Services

Bob Neiers
ARRT Test Specialist

Jerry B. Reid
ARRT Executive Director

Editing, Design, Production Coordination by
Kathy Hendricks, ARRT Director of Communication and Public Relations

published by

The American Registry of Radiologic Technologists®

1255 Northland Drive
St. Paul, Minnesota 55120-1155
Telephone (651) 687-0048
Visit our web site at www.arrt.org

ARRT is the **largest credentialing agency** for radiologic technologists. More than 230,000 individuals hold registration in ARRT's **13 certification categories**. Radiography, nuclear medicine technology and radiation therapy are the primary categories. Post-primary categories include cardiovascular-interventional technology, mammography, computed tomography, magnetic resonance imaging, quality management, sonography, bone densitometry, vascular sonography, cardiac-interventional technology and vascular-interventional technology. ARRT will introduce its breast ultrasound credential in 2004.

Built on three pillars of certification — education, ethics and examination — ARRT's credential is unsurpassed in quality as well as quantity. Indeed, credentials from The American Registry of Radiologic Technologists are referred to as the **gold standard** across the profession.

Educational requirements cover the clinical and didactic approaches, and they continue throughout a Registered Technologist's career. **Ethics standards** further help to distinguish R.T.s and protect patients. "Examination" is the third pillar that rounds out this equation for excellence. And ARRT develops and offers a **state-of-the-art testing** program that is second to none.

How do we do it? By applying the most **sophisticated psychometric concepts** to the development of exam questions, called "items" in the testing arena. Ph.D.-level psychometricians on the ARRT staff work with consultants throughout the profession to create tests that truly measure the **knowledge and skills** required on-the-job. ARRT exam content is rooted in comprehensive practice analyses that hold candidates up to **real-world** levels of expertise.

ARRT's attention to detail is perhaps best exemplified by its item-development process. Test items are designed into proven formats that measure the **critical thinking** and **clinical problem-solving** skills that often elude less-skilled testing agencies.

A certification exam is the pinnacle of a **learning experience** that begins much earlier and continues far beyond the actual test. Today's R.T.s were once challenged by tests offered in their educational programs. Continuing education programs measure their knowledge at every step along the way.

That's why ARRT is sharing its **psychometric expertise** with you. As a team — from preparatory education programs to CE providers — we can help to **ensure high standards of patient care.**

Table of Contents

I) Introduction

Overview 4
The Purpose of Assessment 5
ARRT Exam Content, Length 6
Why Multiple-Choice Format? 7

II) Common Formats for Multiple-Choice Questions

Anatomy of an Item 9
ARRT Item Formats 10
Direct Question, Incomplete Statement, Exhibits, Negatively Worded Items, Combined Response
Formats Not Used by ARRT 17
Multiple True-False, Matching, Comparison, Item Sets, Short Answer, Sentence Completion, Essay, Other
Summary 21

III) Guidelines for item development

General Guidelines 24
Writing the Stem 25
Writing the Distractors 27
Chapter Summary, Item Review Checklist 31

IV) Strategies for Item Production

Five-Step Process 33
Putting It All Together 40
Chapter Summary 42

V) Critical Thinking and Clinical Problem Solving Skills

Levels of Cognitive Complexity 44
Two Important Caveats 45
The Socratic Dialogue 45
Asking 'Why?'...Asking 'What If?'...Asking 'How?'
Assessing Clinical Skills 48
Summary 52

V) Appendices

A. The Length of Certification Exams 53
B. Item Review and Critique 57
C. Evolution of an Item...Into a Test 62
D. Psychometrics Style Sheet 67
E. Bibliography 77

I) Introduction

Some might wonder: “What’s the big deal about writing multiple-choice questions, anyway?” Just by having picked up this manual, you’ve demonstrated awareness that there are probably *several* big deals. By reading it, you will learn what they are and how to manage them.

Overview

This book presents guidelines for developing test items for ARRT examinations. While its primary audience is individuals who write questions for ARRT exams, the information may be useful to anyone who writes test questions, especially those who teach in programs for the health professions. Chapters II and III cover the essentials of item writing: question format, editorial conventions, and stylistic considerations. Chapters IV and V focus on strategies for producing test items. You'll learn how to identify topics…and then convert your knowledge of those topics into test items.

We believe that the best way to learn is by example. So this guidebook is peppered with sample test questions and other exhibits. Most examples are drawn from the field of radiography, but you'll also find occasional examples from other areas of radiologic technology and medical imaging.

The Purpose of Assessment

Testing has many purposes: At the level of the individual student, it is used for evaluating student progress toward achieving educational goals, identifying areas that require remedial activity, assigning grades, and identifying scholarship recipients. At the institutional level, testing is a tool for program evaluation and identifying strong or weak areas of curriculum. At the state or national level, tests are often used to certify competence or to license individuals for practice in an occupation or profession.

ARRT exams fall into this last category.

> The purpose of ARRT certification exams is to:
> *Assess the knowledge and skills underlying the intelligent performance of the tasks typically required by professional practice in the modality.*

Let's look closer at three key phrases:

- ***Assess the knowledge and skills***

  Written exams are great for evaluating cognitive skills, but they don't tell us much about the clinical skills that a person performs in the work setting — positioning a patient, for example. What written exams *can* evaluate is knowledge *related to* positioning: what the position is called, what it's used for, what it looks like, and so on. If you see this as a limitation of written exams, you're right! In fact, this very limitation is one of the reasons behind ARRT's clinical competency and experience requirements.

- ***Underlying the intelligent performance***

  Some test items require candidates simply to recognize facts and basic concepts. But others go deeper — requiring students to understand why certain things are done, how to apply principles to clinical practice, and how to adapt to unusual circumstances. Only the most carefully crafted items can get at these types of critical thinking skills.

- ***Of the tasks typically required***

  ARRT exams are related to practice. They have to be…by law. If an activity or procedure isn't required in practice, then the related knowledge shouldn't be covered on the exam. This is why many of the topics covered in common textbooks aren't addressed by ARRT exams.

ARRT Exam Content, Length

Each ARRT certification exam is built on a task inventory and detailed topic outline. The task inventory is determined by surveying a national sample of practitioners to determine what constitutes typical practice. The topic outline (also known as "content specifications") identifies the knowledge required to perform the tasks and serves as a blueprint for developing the exam. Although ARRT exams and educational curricula cover many of the same subjects, certification exams are more focused on actual practice; so they don't cover *everything* that an educational program does.

The number of questions on ARRT exams varies considerably by area of practice. Most well constructed exams are developed according to the "domain sampling model." No single exam covers all topics in a field; rather, they represent a sample of those topics. Studies have shown that a candidate taking a test with 150 to 200 questions gets a score almost identical to what he or she would obtain on a more exhaustive test of 300 questions. (Appendix A presents a thorough discussion — along with an experiment — that demonstrates this point. Like most appendices, it's optional; read it only if you're really interested in the topic.)

ARRT's post-primary exams typically have fewer items than primary exams because candidates have already demonstrated much of their knowledge. For example, the Mammography exam doesn't cover topics such as radiation protection and basic patient care, because they were already covered on the prerequisite Radiography exam. Some exams are shorter than others because scopes of practice are much more focused. Exams in areas such as bone densitometry require fewer items than exams in MRI or quality management.

Why Multiple-Choice Format?

Although this guidebook focuses on item development for ARRT exams, it may also be useful to instructors, CE providers, and others involved with assessment. So we've included item formats — like matching and true–false — that ARRT doesn't use. Those formats are clearly identified as optional.

ARRT exams use multiple-choice questions (MCQs). These are also known as *selected-response* items, because candidates choose an answer from several that are provided. Essay, short-answer, and practical exams, on the other hand, are referred to as *constructed-respons*e tests. Candidates must actually *create* the answer.

MCQs have had their share of controversy in recent years. Granted, they have their limitations. But they're outweighed by some important benefits (see table).

MCQ: Pros and Cons

Advantages	Disadvantages
• Allow for broad content coverage: many items can be presented and answered in a short period of time	• Test for recognition instead of recall or the ability to arrive at answer without cues
• Best for knowledge and cognitive skills	• Can't assess practical (clinical) or interpersonal skills
• Easy to administer to large groups	• Encourage MCQ study habits like rote memorization
• Can compute statistics to help evaluate item quality	• Scores can be contaminated by "testwiseness" and general reading skills
• Objective scoring: scores are very reliable	

Some of these limitations are just a fact of life, and MCQs aren't always the best assessment option. For example, since we know we can't assess practical skills with MCQs, the ARRT relies on clinical instructors or program directors who observe students in the practice setting.

But most of the limitations to MCQs can be addressed. It's possible to write items that assess in-depth knowledge. It's possible to have technically accurate items that are clearly written at the appropriate reading level. And it's possible to write test items without conveying extraneous clues that benefit testwise candidates.

That's the purpose of this guidebook: to help item writers capitalize on the benefits of MCQs, while minimizing the limitations.

II) Common Formats for Multiple-Choice Questions

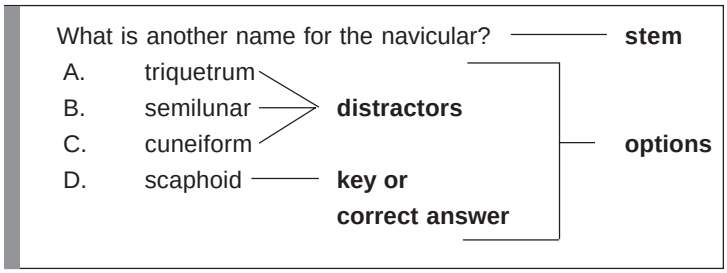
Multiple-choice questions (MCQs) come in all sizes and shapes. Some may have two choices while others have five. Some use “all of the above,” others “none of the above.” Some end in a colon, others end with a question mark.

Indeed, there are many choices to make when writing MCQs. This chapter presents formats that have worked well over the years in a variety of subjects. We start by covering some terminology — just to be sure we’re all using the same jargon. Then we describe several MCQ formats.

Anatomy of an Item

MCQs consist of two parts:

- the *stem*, which presents a problem situation, and
- the *alternatives*, *options*, or *choices*, which provide possible answers.



Stems may be in the form of a question or an incomplete statement. Alternatives include the correct answer (or *key*) and several plausible wrong answers, called *distractors*. The basic format illustrated at left can take on many variations.

The table below summarizes the formats covered in this chapter. Some aren't appropriate for ARRT's certification exams, but are included here because they may be useful for other assessment purposes.

Used on ARRT Exams	**Not Used on ARRT Exams**
Direct Question	Multiple True-False
Incomplete Statement	Matching
Exhibits *(radiographs, tables, illustrations, graphs, etc.)*	Comparison
	Item Sets
Negatively-Worded Items	Short-Answer and Essay
Combined Response	Other: *fill-in-the-blank, true-false, none of the above, all of the above*

ARRT Item Formats

Two basic MCQ formats comprise the majority of test items on ARRT's certification examinations: direct question and incomplete statement. Adding tables, graphics, and other material can enhance both, as described later.

Direct Question

In the example at right, the stem poses a single, complete problem, while each option provides a possible solution or answer. Item writers prefer this format for its ability to present a clearly formulated problem.

> Which of the following refers to the degree of blackening seen on a radiograph?
> A. radiographic intensity
> B. radiographic contrast
> C. radiographic sharpness
> D. radiographic density *
> *(* Asterisks in examples indicate correct answers)*

MCQs typically consist of three, four, or five choices, or options. The more choices candidates have, the less likely they can just guess the correct answer. It's often difficult, though, to create five plausible choices; and adding obvious wrong answers doesn't increase an item's quality or difficulty. In fact, some experts recommend just three options. For our purposes, we will concentrate on MCQs with four options.

Incomplete Statement

The preceding example can also be written as an incomplete statement, a format that often results in fewer words. In this case, the word *radiographic* was deleted from each option and placed into the stem.

> The degree of blackening seen on a radiograph is referred to as radiographic:
> A. intensity
> B. contrast
> C. sharpness
> D. density *

Item writers occasionally confuse the incomplete statement and sentence completion (fill-in-the blank) formats. For example, an item might appear as: "The degree of _____ that is seen on a radiograph is referred to as density." The candidate's task is to identify the word or phrase that goes in the blank (e.g., "blackening" or "darkness"). Most experts agree that the sentence-completion format often leads to items that aren't very clear. ARRT doesn't use this format, and we suggest that you avoid it — even for classroom tests.

FLAWED ITEM: UNFOCUSED STEM

The degree of blackening seen on a radiograph:

A. must be controlled by the kVp setting
B. is called contrast
C. is frequently caused by scattered radiation
D. is referred to as radiographic density *

The key to writing a good incomplete statement is to have the stem convey a complete problem. It might be tempting to break a statement by inserting a colon (:) into the middle of it, and then to call it a stem. But you can end up with an item that's flawed, like the example at left.

TIP:

A good way to tell if a stem is unfocused is to mask the options and ask whether a knowledgeable candidate could determine the answer by reading just the stem. If it's necessary to read all of the options to figure out what the stem is really asking, then the stem doesn't have ample information.

An even better way to focus a stem is to first write it as a direct question, and then *convert* it to an incomplete statement — but only if there's good reason to do so, such as making the item easier to read.

In general, if it's not possible to phrase a stem as a question, then the item is probably flawed.

This example of an "unfocused stem" demonstrates the most common problem in item writing: not only does the stem fail to present a single problem, it also — because it lacks focus — encourages heterogeneous options, which tend to wander all over the place. Worse yet, a stem that doesn't present a well-defined problem makes it easier to end up with two correct answers (e.g., option C is partially correct because scatter can increase density).

Note that the previous examples contain only one correct answer, and the distractors are clearly incorrect. Not so with a variation of the direct question and incomplete statement formats called "best answer." It offers alternatives that may be partially correct, with one being clearly *more* correct than the others. The best-answer format is used to gauge complex achievement, such as asking a candidate to select the best reason for an action, the best method for doing something, or the best application of a principle. For example, an item on managing a skin reaction to radiation therapy might list several acceptable ways to treat erythema, but one of the methods may be better than the others. Best-answer items usually require judgment or opinion; they should be used judiciously.

Exhibits

The incomplete-statement and direct-question formats serve as building blocks for items that display information in exhibits such as tables, graphs, text paragraphs, drawings, medical images, and photographs. Exhibits can present information in a way that's practical, clinically relevant, and concise ("a picture says a thousand words"). They also provide an alternative way to assess critical thinking.

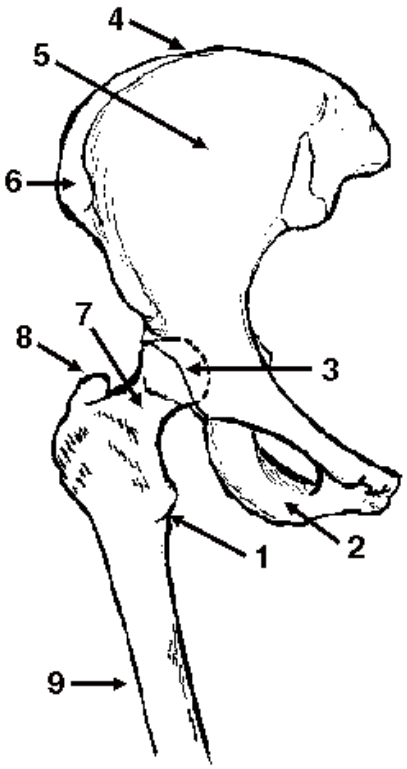
Given that radiologic science is mostly about medical imaging, it makes perfect sense to use radiographs, CT scans, sonograms, and other images for test items. On most ARRT exams, 10% or more of the test items present a medical image or illustration. And once you've gone to the trouble of producing an image or illustration, it makes sense to write several items to go with it. The following examples demonstrate the different types of items that can accompany an image or illustration.

The first two items in the box below cover basic anatomy. Although nearly identical, they ask the question in distinctly different ways. Which item is easier? We suspect the first one: it supplies more information by giving more anatomical labels — like providing more pieces of a puzzle.

The remaining examples show that drawings can go beyond simple identification of parts. The third example in the set addresses physiology, while the last two items require detailed knowledge of specific radiologic studies.

Keep in mind that most illustrations lend themselves to questions about both structure (e.g., anatomy) and function (e.g., physiology). For example, test items based on a diagram of a film processor could require candidates to identify parts, state the materials they consist of, explain how they work, or identify where different malfunctions may occur.

TIP:
For a set of items related to the same exhibit, options should use *either* all words (like the first in the box below) *or* all numbers (remaining items). Mixing them up, as we did here, may result in items that "clue" one another.



Arrow 4 in the illustration on the left identifies the:
A. iliac crest *
B. ischium
C. greater trochanter
D. femoral head

The iliac crest is identified by arrow number:
A. 1
B. 2
C. 3
D. 4 *

The rate of bone turnover is usually highest at which of the following regions?
A. 2
B. 3
C. 5
D. 7 *

Which arrow identifies a site commonly used to measure bone mineral density?
A. 4
B. 5
C. 7 *
D. 9

The modified Cleaves position (or bilateral frog leg) demonstrates what anatomical part?
A. 2
B. 6
C. 7 *
D. 9

A bone densitometry scan of the lumbar spine produces the BMD measurements below. What might explain these results?

region	g/cm^2
L1	0.98
L2	0.92
L3	0.96
L4	1.01

A. osteoporosis
B. scoliosis
C. vertebral fracture at L1 *
D. severe facet sclerosis at L2

Which of the following sets of technical factors will produce a radiograph with the greatest density?

	mA	msec	SID	kVp
A.	100	500	48”	80
B.	200	250	36”	76 *
C.	400	125	40”	76
D.	500	100	40”	76

Tables and graphs are other common types of exhibits. Even a list of words or paragraph of text might be considered an exhibit. The first example at left highlights the advantages of using a table to organize information in the stem. Although this table provides numbers, tables can also be used to present text, such as a list. (An item in Chapter 5 presents a list outlining the steps of a QC procedure.) The second example presents a table in the options. While not technically an exhibit, the table format does improve readability.

Exhibits — such as images, illustrations, tables, graphs, and other types of information displays — can be used as the basis for numerous types of test items. The table below summarizes some of the possibilities.

Type of Exhibit	Knowledge and Skill Assessed
Medical images, anatomical illustrations, positioning photographs or diagrams	Recognize anatomy, physiology, or pathology; identify positions and projections; recognize positioning errors; identify artifacts; evaluate image quality; compensate for poor images
Drawings or photographs of equipment and instrumentation	Recognize parts, explain how systems function, explain QC procedures, troubleshoot equipment, evaluate instrument settings
Models of scientific principles or processes (e.g., x rays interacting with a molecule, dose-response curves)	Identify and label parts, infer relationships, determine words or equations that correspond to a model or curve
Tables or graphs with technical factors, technique charts, equipment specifications, results of QC tests	Evaluate technical factors, calculate certain results, interpret and evaluate QC data, draw inferences, interpret data and use charts (e.g., bone density results; D log E curves)

Negatively Worded Items

You've seen test items that use *not*, *except*, or other negative wording in the stem. Sometimes it makes perfect sense to write test items that emphasize what should not be done — such as when certain drugs or procedures are contraindicated, or when some action must be avoided because it could be harmful. But negative wording can result in flawed items, so ARRT uses it sparingly.

The first example at right was retired from the Radiography exam. As we shall see, it's not a very good item and probably could have been written in standard MCQ format with just one correct answer. It was probably written this way because: (1) grid cut-off has different causes; and (2) the item writer probably had difficulty coming up with three plausible, but incorrect, answers.

All of the following will result in grid cut-off EXCEPT:

A. an off-center tube
B. a tube that is perpendicular to the lead strips
C. improper SID being used with a focused grid
D. grid motion being started before exposure is made *

Which of the following will NOT result in grid cut-off?

A. an off-center tube
B. a tube that is perpendicular to the lead strips
C. improper SID being used with a focused grid
D. grid motion being started before exposure is made *

The *except* format is marginally acceptable, but it's difficult to write options that are grammatically consistent with the stem. The second example above illustrates a better way to state negatively worded items. It's easier to read and more likely to result in consistency.

A major limitation of these examples is that quick readers may miss the negative phrasing and choose an incorrect answer. To help prevent this type of oversight, always highlight the negative word in uppercase, bold, or italic text (NOT, **except**, *unless*). Another limitation is the tendency to end up with double negatives. Option C in the preceding examples contains the negative *improper*, which makes the items difficult to understand clearly.

Which of the following may result in grid cut-off?

1. an off-center tube
2. grid motion being started before exposure is made
3. improper SID being used with a focused grid

A. 1 & 2 only
B. 1 & 3 only *
C. 2 & 3 only
D. 1, 2, & 3

Which of the following may result in grid cut-off?

1. an off-center tube
2. grid motion being started before exposure is made
3. increased OID

A. 1 only *
B. 2 only
C. 3 only
D. 1, 2, & 3

Which of the following interactions between the x-ray beam and human tissue typically affect the quality of a radiograph?

1. photodisintegration
2. Compton effect
3. photoelectric effect

A. 1 & 2 only
B. 1 & 3 only
C. 2 & 3 only *
D. 1, 2, & 3

Combined Response

This format is a close relative of the negatively worded item. A combined-response item (also called the *type k* format) consists of a stem followed by three or four possible responses, one or more of which is correct. The first example at left is similar to the negatively worded items already presented in that two of the responses are true (or correct), while one is not. In the second example, only one of the responses is correct.

The best use for the combined-response format is when multiple things are correct, and candidates should be able to consider them simultaneously. For example, when considering: (1) how different technical factors affect contrast, density, or other imaging factors; (2) how multiple symptoms must be present at the same time in order to make a diagnosis; (3) how different events jointly interact to influence an outcome. The third example at left illustrates an effective use of the combined-response format.

Some educators mistakenly believe that combined-response items assess critical thinking. Although this format may require more mental processing, it's the type of skill that's easily taught in "test coaching" classes. The following example illustrates how clever candidates who know little about a topic can work their way through combined-response items that aren't carefully written.

Testwise Rule 1: Look for the responses that appear most often.

Every coaching school teaches this trick — called convergence strategy. Response 1 appears three times. Response 3 also appears three times. Since option B contains both 1 and 3, it could be the correct answer.

Testwise Rule 2: Find a response you know to be incorrect.

Since response 4 won't fool too many, a half-knowledgeable candidate will rule out both C and D.

FLAWED ITEM: COMBINED RESPONSE

Which of the following may result in grid cut-off?

1. an off-center tube
2. grid motion being started before exposure is made
3. improper SID being used with a focused grid
4. increased OID

A. 1 & 2 only
B. 1 & 3 only *
C. 1, 3, & 4 only
D. 2, 3, & 4 only

Testwise Rule 3: Find a response you know to be correct.

If the only thing a candidate knows about grid cut-off is that it can occur when the tube is off center, he or she can narrow the answer down to A, B, or C. (The strategy helps very little in this particular case.)

Testwise Rule 4: Look for any other obvious clues.

Response 3 has the word *improper*, which is a pretty good clue that something bad could happen — like grid cut-off. This suggests that B, C, or D could be correct.

To recap, at least four paths can lead to a correct answer on combined-response items:

- what you know to be correct
- what you know to be incorrect
- obvious clues
- convergence strategy (option counting)

Although it's unlikely that someone would apply all the strategies to a single item, using any two of them greatly increases the chances of getting the item correct with just partial knowledge.

Some of those potential giveaways are avoidable. For example, using each response an equal number of times can minimize the success of convergence strategy.

Because negatively worded and combined-response items are susceptible to so many flaws, the ARRT will not accept any more than 10% of items in this format from an item writer. If you write them, do so sparingly and with caution.

Formats Not Used by ARRT

ARRT doesn't use the following formats, but they can be effective in educational settings. One advantage is that they provide a way to assess specific topics in great detail. However, this can also be a limitation: if much testing time is devoted to one topic, then perhaps too little time will be left for others. Another benefit to these formats: they provide instructors with more detailed information for diagnosing a student's strengths and weaknesses.

Multiple True-False

This format is a distant relative of combined-response and negatively worded items. It has an edge over combined-response in that it minimizes testwiseness. Because each response is scored separately, a single item can zero in on the exact part of a concept that troubles the student, which is difficult to do with combined response.

Grid cut-off may result from:

an off-center tube	(T* F)
grid motion being started before exposure is made	(T F*)
improper SID being used with a focused grid	(T* F)
increased OID	(T F*)

For each of the following statements about grids, indicate if it is true or false.

The use of grids improves radiographic contrast.	(T* F)
Grids should be used routinely for low kVp procedures	(T F*)
Grids reduce scatter	(T* F)

A multiple true-false item consists of an introductory phrase or stem, followed by a set of specific items related to the stem. The student indicates whether each item is or is not true. This format can also be helpful as a starting point for writing several items in the standard MCQ format (i.e., four options with only one being correct).

The multiple true-false format is a breeze for writing items that assess simple recall of facts. And, as with standalone true-false items, it is possible to craft complex problem-solving items; it just takes a little extra effort.

Matching

Although matching items are seldom used on certification exams, they do have some legitimate uses, especially in educational settings. They are most effective for assuring that students understand similarities and differences among closely related concepts. Used in conjunction with an illustration or image, they can also be useful for assessing anatomy and physiology knowledge.

The following examples provide instructions, an answer list, and several items. Five to ten items is most efficient.

Directions: Appearing below is an answer list followed by several statements. For each statement, choose the one answer (A, B, C, or D) that is most closely related to the statement. Answers may be used once, more than once, or not at all.

Answer list

A. radiographic density
B. radiographic contrast
C. recorded detail
D. distortion

Statements

Usually increases with increasing magnification	(D)
Degree of blackening of a radiograph	(A)
Inversely proportional to kVp	(B)
Also known as resolution	(C)
Controlled by mAs; influenced by kVp	(A)
A misrepresentation of size or shape	(D)

Directions: The following items require you to compare related concepts. Appearing below is an answer list followed by several statements. For each statement, choose the one answer (A, B, C, or D) that is most closely related to the statement. Answers may be used once, more than once, or not at all.

Answer list

A. radiographic density
B. radiographic contrast
C. both A and B
D. neither A nor B

Statements

Degree of blackening of a radiograph	(A)
Inversely proportional to kVp	(B)
Can be controlled by radiographer	(C)
Also known as resolution	(D)

Comparison

Best reserved for determining whether students understand closely related concepts, comparison items are very similar to matching items and have the same limitations. Note in the item above that option C is similar to *all of the above*, while option D corresponds to *none of the above*. This format can be a little tricky, so don't overuse it. Again, it may be desirable to have several items for each answer list.

A 43-year-old male is referred for a BMD study by his primary care physician. His records indicate he received a heart-lung transplant 17 months ago. While giving his history, the patient mentions that, other than becoming easily fatigued, he has generally felt well in recent weeks. He indicates that he has been taking prednisone for the past 12 months as prescribed by his physician. He has no history of fractures. This is his first BMD examination.

Will this patient's bone scan be covered by Medicare?

A. No; there is no prior evidence of prior fracture or bone demineralization
B. No; Medicare reimbursement is provided only for post-menopausal women
C. Yes; Medicare covers patients receiving glucocorticoid therapy *
D. Yes; Medicare covers any patient referred by a physician

Which of the following scans would be best for establishing and monitoring this patient's BMD?

A. DXA scan of the forearm
B. DXA scan of the hip *
C. RA scan of the finger
D. ultrasound scan of the heel

A scan of the spine (L1 to L4) produces the values below for bone mineral content and total area scanned. What is the total BMD for L1 to L4?

BMC (L1 – L4) = 56.2 g
Area (L1 – L4) = 58.4 cm^2

A. 0.16
B 0.96 *
C. 1.04
D. 2.20

About how often should this patient receive follow-up scans?

A. weekly
B. every 4 to 6 weeks
C. every 6 to 12 months
D. every 5 years

Item Sets

The purpose of an item set is to simulate the clinical decision-making process that occurs in the practice setting. An item set consists of a paragraph or exhibit followed by several test items that refer back to that information. Item sets are most suitable for assessing interdependent sets of skills, especially procedures that require multiple steps or decision points (e.g., obtaining multiple mammograms of the same patient, evaluating the results of a QC procedure and taking corrective action, providing radiation therapy to a patient over time). In the examples at left, all items relate to a patient referred for bone densitometry.

Ideally, all items should build on information provided in the initial paragraph. Although each item above relates to the paragraph, not all of them actually require the information it provides. For example, the third item would function adequately as a stand-alone question. This is probably a pardonable sin for one or two items, but if most of the items can stand alone, then why even present them as a set?

Watch for three potential limitations with item sets. First, *clueing* is more likely, because one item may provide the answer to another. Second, a student who overlooks a single piece of information might get nailed on not just one item, but several. Third, although each item should relate to the initial paragraph, they should be independent of each other, i.e., knowing the second item shouldn't depend on knowing the first.

Item sets provide an excellent way to assess clinical problem-solving skills. In fact, ARRT previously used them on selected exams. Once computer-based testing came along, item sets were out of the picture. The computer screen's limited size couldn't display all of the relevant information.

Short Answer, Sentence Completion, Essay

These formats, although not multiple choice, merit brief mention because they can be useful in classroom settings. But they're time-consuming to score, and are not practical for most large-scale testing programs. Here are two examples of the short-answer format.

> What term is used to describe the degree of blackening seen on a radiograph?
>
> The degree of blackening seen on a radiograph is referred to as _____ .

These items are similar to the two presented at the beginning of this chapter, the major difference being that they have no options. And, being optionless, it's imperative that the stem completely state the problem, which the direct question is just better at it. All too often, the incomplete statement ends up being a "guess what the teacher is thinking" exercise.

The short-answer format can be especially useful for professional item-writers and educators whose ultimate goal is to write standard MCQs. As any experienced item writer knows, the hard part is coming up with distractors. If a teacher includes a few short-answer questions on every exam, the incorrect answers that the students generate could make mighty fine distractors for future MCQs.

Short-answer format is closely related to a variation that we don't recommend: sentence completion, or fill-in-the-blank. It's the blank that is the problem: there's no good place for it. Don't even think of putting it at the beginning or middle of the item, because that can result in a vague stem that leaves the candidate trying to figure out what the item is even asking. Which leaves the end. And an item with a blank at the end actually ends up as the incomplete statement format we've already examined.

Good test items should require candidates to solve problems, not guess what the item writer is thinking. Although the fill-in-the-blank format is perfectly acceptable for reading comprehension exams, they are seldom useful for professional certification exams.

The essay question, an extension of the short-answer format, is good at gauging the depth of a student's knowledge about a particular topic. To encourage focused responses, essay questions must be clearly stated and provide ample direction to the student. Because grading is very subjective, establish criteria to help assure fair and reliable scores.

Other Formats

A few other widely used formats are worthy of brief mention — primarily to discourage their use.

- Avoid formats such as "none of the above" and "all of the above." Anything they may accomplish can be better achieved with the combined-response format.
- Standalone true-false items have some merit, primarily because they allow more questions to be asked in a shorter time frame. But building a good true-false exam can be a challenging task.

Summary

This chapter described several useful — and some not-so-useful — formats for assessing knowledge and skills in the radiologic sciences and other disciplines. The table below lists each format and offers a summary judgment. The next chapter discusses style and editorial guidelines for those formats that ARRT endorses.

Review of Item Formats

ARRT-Endorsed

Direct Question	Preferred format
Incomplete Statement	Also acceptable
Exhibits *(illustrations, medical images, tables, graphs, etc.)*	Very desirable
Negatively Worded	Acceptable in limited numbers (flaws)
Combined Response	Acceptable in limited numbers (flaws)

Not Endorsed

Multiple True-False	Beneficial for some applications
Matching	Beneficial for some applications
Comparison	Similar to, but less useful than, matching
Item Sets	Very effective
Short-Answer and Essay	Useful in classroom or small group settings
Others *(fill-in-the-blank, true-false, none of the above, all of the above)*	Avoid these; susceptible to flaws

III) Guidelines for Item Development

The previous chapter focused on MCQ formats. This chapter delves into another important aspect of item writing: style. Guidelines from this chapter will help you produce items that are consistent in both format and style. While we urge you to keep these guidelines in mind, we're quick to point out that they are just that — guidelines, not rules. Some are stylistic choices made by ARRT staff; others represent recent philosophy in teaching and assessment. General adherence to these guidelines will result in well written, useful items that won't require additional editing.

This chapter is divided into three major sections. First we cover some general guidelines for item writing. Then we get into the components: the stem (which presents the question or problem) and the options (distractors as well as the correct answer). The chapter summary pulls it all together into a handy checklist.

Before we start, a rule of thumb: MCQs usually present three, four, or five options. The more choices, the less likelihood of a guess being right. In theory, with five options there's only one chance in five of guessing the correct answer; with four options, it's one chance in four. But it can be difficult to come up with five plausible choices. And neither item quality nor difficulty is discernibly increased by adding obviously wrong answers that no one would choose. In fact, studies show three or fours options per item to be optimal. For our purposes, we'll concentrate on creating MCQs with four options.

General Guidelines

Here are seven standards — plus a big tip — for item writers:

1. Test important knowledge and skills. Design each item to measure an important learning outcome. Keep the focus on the content of the item; resist the temptation to include irrelevant, obscure, or insignificant material in an attempt to increase the item's difficulty. (Chapter IV offers strategies for identifying specific topics.)

2. Be up-to-date and correct. Avoid referring to events in the immediate future; it all-too-soon becomes the past. Try to write items that are contemporary, but won't appear dated in a year.

3. Provide sufficient information. The stem of an item must contain sufficient information to enable the candidate to select the appropriate answer. The goal is to test candidates' knowledge, not discover how well they can take a test.

4. Cite "according to whom." If accepted authorities don't agree on the best response, select and cite the appropriate one (e.g., "According to ACS guidelines…"). The NCRP recommendation on radiation protection may differ from the regulations in your state. Both are right, but you have to let the candidate know which one the test item is about. Because ARRT exams are national in scope, we almost always rely on national authorities rather than state or local ones.

5. Make distractors plausible. All distractors should be logically consistent with the stem. Any that are silly or obviously incorrect will compromise an item's validity.

6. Avoid bias. Items must avoid both the reality and appearance of bias, in terms of gender, culture, race, and other discriminatory factors.

7. Offer only one correct answer. It almost goes without saying: there can be one and only one correct, or clearly best, answer.

Item writing is a two-part task: every item has a stem and four options. It doesn't matter whether you start with the correct answer and work back to an appropriate stem, or vice versa — these guidelines will gear you toward success.

TIP:

- State each item clearly and grammatically.
- Write simply: no overly long, complex, confusing sentences.
- Make any pronoun referents unmistakable.
- Avoid slang, and stay away from words, phrases, or constructions that make reading unnecessarily difficult.
- Direct questions make better stems than incomplete sentences do. And they help avoid other problems such as incomplete stems, non-parallel options, and convoluted wording.
- When an incomplete sentence is unavoidable, be sure to complete it in the appropriate syntax.

Writing the Stem

1. Set the task

A stem must provide sufficient information for the candidate to be able to interpret the item's intent and select the appropriate answer. Candidates shouldn't have to read all the options to figure out what you're asking for. Check the clarity and completeness of the stem by covering the alternatives and determining whether the item could be answered without them.

2. Make it clear

Avoid ambiguous, confusing, or vague wording. The only reason to include superfluous material would be if you were testing whether candidates can identify relevant information. Test items are supposed to allow candidates to show what they have learned. If an item's wording, vocabulary, or sentence structure is confusing and prevents candidates from understanding what they're being asked, they won't have that opportunity.

UNFOCUSED STEM

According to the NCRP, the occupational dose-equivalent limit to a pregnant radiographer should NOT exceed how many rem during the gestation period?
A. 0.3 rem
B. 0.5 rem
C. 1.0 rem
D. 5.0 rem

The _____ common side effect of treatment with _____ is myelosuppression.
A. most/radiation
B. least/radiation
C. most/chemotherapy
D. least/chemotherapy

The primary purpose of the x-ray performance standards specified by Title 21 of the Code of Federal Regulations (part 1020) is to regulate:
A. maximum patient exposure to x rays
B. the design and manufacture of x-ray systems
C. radiologic terminology
D. radiographer knowledge and skill requirements

The examples at left share one fundamental fault: they're fuzzy, due to awkward sentence structure, confusing wording, or vague terms.

The first item's unusual sentence construction plus use of the negative equals confusion. This item is testing relatively basic information, which can best be approached with a simple straightforward question: "According to the NCRP, what is the occupational dose-equivalent limit (in rem) for a pregnant radiographer?"

The second example is practically incomprehensible. Candidates will get it wrong — not because they lack the knowledge, but because they can't figure out what the actual question is. Stems must be clearly and simply worded to set the task. Candidates should know exactly what they're supposed to do or what type of information is being sought. The "fill in the blank" format often results in unclear stems, which is one reason why ARRT exams don't use them. Here is a much better version of the same item: "What is the most common side effect of treatment with chemotherapy?"

An ambiguous term in the third example on the preceding page is likely to cause candidates to stumble. What does the word *primary* mean? Couldn't one make a case for each one being "primary" in some way? Words like *best, worst, most important,* and *greatest* must be clarified. Supply additional information that answers the question, "In terms of what?"

3. Target appropriate level of reading difficulty

Item difficulty is dramatically influenced by vocabulary level. Even a simple idea can be encased in vocabulary that very few candidates would understand. Notice how the following three examples demonstrate — the first two, rather egregiously — a test of vocabulary, rather than subject matter:

- The postulation of capillary effectuation promotes elucidation of the mechanism by which pliant substances ascend in incommodious veins.

- The thesis of capillary execution serves to illuminate how fluids are elevated in small tubes.

- The principle of capillary action helps explain how liquids rise in small passages.

Clearly, vocabulary level — as well as sentence structure — has a major influence on whether candidates will understand what's being asked in a test question. Simple, declarative sentences work best.

4. Keep it short, but not too….

State the stem as precisely as possible and steer clear of unnecessarily complex wording and sentence structure. A stem should present a complete problem. Examine the item at right.

The stem fails to clearly set a problem. A candidate would have no idea what question is being asked. Only after reading the stem *with* all of the options does its point become clear. Candidates should know as soon as possible exactly *what* is being asked. Here's a rewrite that not only sets the problem clearly, it also relates each option to the intent of the item.

BAD:

The navicular

A. is sometimes used as another name for the scaphoid *
B. can be located in a patient's skull
C. is a small bone found in the knee
D. is a blood vessel

BETTER:

What is another name for the navicular?

A. triquetrum
B. semilunar
C. scaphoid *
D. cuneiform

Writing the Distractors

1. Make one answer the correct one

Multiple-choice items ask candidates to choose a single correct answer from the options provided. Presented with more than one plausible answer, the candidate faces a dilemma — deciding which is the proper one. Candidates shouldn't have to be mind-readers to figure out the intent of a question. That's the item writer's responsibility.

2. Don't give unintended clues

Avoid sending signals that might help candidates select the correct answer or eliminate an incorrect alternative. Most extraneous clues in multiple-choice items are found in the wording. Two key sources are the "implausible option" and the "specific determiner."

Note the grammatical clues in these two examples:

GRAMMATICAL CLUE: ARTICLE

A lateral malleolus is associated with an:

A. knee
B. shoulder
C. ankle *
D. hip

In the first, the common mistake of using the article "a" or "an" at the end of the stem is an important indicator of the correct answer. If a candidate already knows that the correct answer to the item is either "proper" or "improper," the article "a" before the blank indicates that the next word must begin with a consonant, so the candidate will correctly guess "proper." Eliminate this problem by placing the article in the options.

A lateral malleolus is associated with:
A. a knee
B. a shoulder
C. an ankle
D. a hip

GRAMMATICAL CLUE: VERB

An electrical transformer can be used:

A. storing electricity
B. to increase the voltage of alternating current *
C. it converts electrical energy into mechanical energy
D. alternating current is changed to direct current

In the second example, only answer B grammatically fits the stem. Regardless of whether candidates know anything about electrical transformers, this clue will lead them to the answer. Be sure that all distractors match their stems in terms of logic and grammar.

Another common, but perhaps less obvious, mistake is presenting options that are inappropriate or implausible. Candidates can dismiss such distractors immediately and increase their chances of choosing the correct answer.

Which of the following best describes an electron?
A. a negatively charged particle *
B. a neutral particle
C. a positively charged particle
D. a voting machine

Answer D, while clever, is so obviously wrong that no candidate who reads the item carefully will select it. The more plausible choices a candidate has, the less likely that he or she can simply guess the correct answer. Item writers can usually think of two or three good distractors for a stem, but have difficulty coming up with a fourth. Avoid the temptation to use either *all of the above* or *none of the above.* They only confuse the candidate and add nothing to the validity of the item…which leads to the next rule.

3. Avoid "all of the above" and "none of the above"

Using *all of the above* as an option lets a candidate answer an item on the basis of partial information. They can tell that *all of the above* is the correct choice simply by knowing that two of the alternatives are correct. In turn, they can tell that it's wrong by recognizing that at least one of the alternatives is incorrect. Once a candidate has determined that "all of…" and "none of…" are not the correct choices, his or her chances of guessing correctly have doubled.

Using *none of the above* as a correct answer does nothing more than measure the ability to detect incorrect answers. The candidate doesn't demonstrate knowledge of what's correct. This alternative is used most often in computational problems where — if it is the correct answer — the candidate would still have to perform the computation in order to know.

Another clue too often found in the vocabulary of many items is the *specific determiner.* Using terms like *some, sometimes, often, may, always, never, all,* and *none* can tip off answers. It's not so much that the correct answer is given away; rather, it's that candidates — because they know that few things in life "always" or "never" happen — can eliminate distractors.

4. Make structure parallel

Options for any one item should all begin with the same part of speech and be approximately the same length. A correct answer that's noticeably longer or shorter than the distractors draws immediate attention to itself. Longer options are frequently correct; it's their additional detail that makes them correct. Avoid items with a correct answer that's detailed and distractors that are fuzzy, vague, or incomplete.

OUT OF ORDER:
An exposure is obtained at 200 mA for 0.5 seconds. What is the mAs?
A. 1000 mAs
B. 0.25 mAs
C. 400 mAs
D. 100 mAs *

RATHER:
An exposure is obtained at 200 mA for 0.5 seconds. What is the mAs?
A. 0.25 mAs
B. 100 mAs *
C. 400 mAs
D. 1000 mAs

5. Arrange in order

Present options in some logical order. For items that require a numerical response, present the choices in either ascending or descending order. Short verbal responses can be arranged in alphabetical order unless there's a reason to present them in some other manner. Anatomical parts can also be arranged in order (e.g., anterior to posterior, proximal to distal, and so on).

Use an efficient and logical format. Listing options on separate lines, one under another, makes them easy to read and compare. Use letters to differentiate the choices, since numerical answers in numbered options could confuse candidates.

6. Vary the correct answer's position

Place correct answers randomly. And don't feel like you have to use B or C every time. Correct answers should appear in each position roughly the same number of times, but their placement shouldn't follow a discernible pattern. In a 25-item quiz, for example, option A should be used as the correct answer five to seven times, option B five to seven times, and so on.

7. Be careful with opposites

Why — in fields like biology, physics, and healthcare in general — does it seem natural to write items using distractors that are opposites? Because physical events have effects that can occur in one direction or another. For example, a question might ask what happens to contrast as kVp increases. Well, contrast will do one of two things: increase or decrease.

How do we deal with such items? Consider this example:

What should be changed to increase the depth of penetration of an ultrasound beam?
A. decrease frequency *
B. increase frequency
C. increase the velocity
D. decrease the amplitude

Options A and B are related to each other as paired opposites. Meanwhile, options C and D are unrelated. A rule taught in test coaching school: if two of the options form a pair, one of them is likely the correct answer. In this case, the testwise student will immediately eliminate options C and D. Even if that candidate knows nothing, he or she suddenly has a 50% chance of getting it right by guessing.

Instead, write options for this type of item in one of the two ways shown in the box at right.

Form two pairs of opposites, like this:
A. decrease frequency *
B. increase frequency
C. increase velocity
D. decrease velocity

Or, avoid opposites altogether, as shown here:
A. decrease frequency *
B. increase intensity
C. increase velocity
D. decrease amplitude

The first of these items initially requires candidates to recognize that frequency, not velocity, controls penetration. Next they have to know that depth of penetration is increased by decreasing the frequency.

Here's one more example. Although the first set of options might be acceptable, some candidates will quickly rule out option B because it is different. Meanwhile, others may actually choose option B because it is different. And still others will go straight to C or D because they are opposites.

For a PA oblique projection of the optic foramen, the central ray should be directed:
A. parallel with the interpupillary line
B. perpendicular to the interpupillary line
C. parallel with the acanthomeatal line *
D. parallel with the glabellomeatal line

The set of options at right is better. They are parallel in structure (two pairs of opposites) and provide less clueing. Candidates must discern between two lines (interpupillary vs. acanthomeatal) and two planes (parallel vs. perpendicular).

A. parallel with the interpupillary line
B. perpendicular to the interpupillary line
C. parallel with the acanthomeatal line *
D. perpendicular to the acanthomeatal line

8. Avoid cross-keying

Make sure that the content of one item doesn't provide the answer to another. And avoid synonyms or paraphrases in options. Candidates who recognize them will realize that they are, in effect, the same answer — but can't both be the correct choice.

Chapter Summary

As stated at the beginning, these are only guidelines. Follow them when they're logical; break or bend them if doing so will improve an item's effectiveness. The following checklist may be helpful for evaluating MCQs written by you or others.

Item Review Checklist

- Does the stem present a single, clearly formulated problem?
- Is the stem stated in simple, clear language?
- Is the stem worded so there is no repetition of material in the alternatives?
- Is the stem stated in positive form whenever possible (limited use of *not, except*)?
- Is the item written in a manner that's unbiased in terms of gender, culture, race, and other factors?
- If negative wording is used in the stem, is it emphasized either by underlining, bolding, or capitals?
- Is the intended answer correct or clearly best, and is there only one correct answer?
- Are all the options grammatically consistent with the stem?
- Are all the options parallel in structure and form?
- Are the options free from verbal clues to the correct answer?
- Are the distractors plausible to the uninformed?
- To eliminate length as a clue, is the correct answer about as long as one or more of the distractors?
- Have *all of the above* and *none of the above* options been avoided?
- Is the position of the correct answer varied so there's no detectable pattern?
- Are alternatives in logical order when appropriate or in random order when called for?
- In general, does the item's format and grammar facilitate efficient and easy test taking?

IV) Strategies for Item Production

Chapters II and III described several MCQ formats and discussed editorial guidelines for stems and options. That was the easy part. The hard part is sitting down to put your ideas on paper. This chapter describes a five-step process for turning ideas into test items. It's intended to help item writers organize their thoughts and overcome writer's block.

Five-Step Process

For most writers, the task begins by thinking of a topic and doesn't end until after editing what they've written.

We have divided the production process into five steps because it's convenient, seems logical, and bears some similarity to what some of our better item writers have done over the years. Whether you actually use three, four, five, or ten steps to write items is pretty much a matter of personal work style.

The five steps we propose are:

1. Identify a topic and supporting information
2. Determine the candidate's task
3. Formulate the stem and correct answer
4. Write the distractors
5. Review the item

Before getting into the details of each step, a couple of caveats are in order. First, although each step can be described as a separate activity in theory, in practice they meld together to form a more-or-less continuous process. Second, it's not always necessary to perform each step; sometimes they just occur automatically. We still find it helpful to break the process down into discrete steps that can be analyzed and discussed. (Such discussion has the bonus value of sometimes helping overcome writer's block.)

1. Identify the topic and supporting information

Most exams are assembled according to a blueprint called *test specifications*, and this is true of ARRT's exams. The content specifications consist of a topic outline and a task inventory. The topic outline identifies the knowledge to be tested, while the task inventory indicates how that knowledge may be applied in the practice setting. Both pieces can be helpful for item writing.

To demonstrate how the topics and tasks work together, we'd like to work through an example. Consider this task from the Radiography task inventory:

> *Modify exposure factors for circumstances such as involuntary motion, casts and splints, pathological conditions, or patient's inability to cooperate.*

Although the physical task of modifying exposure factors is fairly simple, doing it correctly requires extensive knowledge of numerous topics including attenuation, the relationship between mAs and penetration, impact of pathological conditions on tissue density, and so on. You recall from your own experience that this is indeed an important task and that, as an entry-level technologist, your knowledge of related subjects — especially attenuation — was critical to performing this task effectively. So you decide to write an item on attenuation.

Attenuation is a basic subject on which it would be fairly easy to write a test item without any references. But don't be so quick to pounce! Consulting references or other sources can be helpful in several ways:

- Test specifications often lack the detail needed for item development. A good textbook or article will fill in any gaps.
- References can verify that you're using precise terminology (e.g., is the correct term *technical factors* or *exposure factors*? *Compton scatter* or *scattering?*).
- Looking into a reference will likely lead you to ideas for several test items instead of just one or two.

TIP:
Consulting references not only makes for better items, it also leads you to additional items you may not have otherwise considered.

We suggest that item writers refer to textbooks, major reports (e.g., NCRP, NRC, OSHA), curriculum guides, or lists of learning objectives, also called behavioral objectives. Major textbooks are probably the single best source of supporting information.

So, to continue with our example, let's dig out an old textbook and look at the chapter on attenuation. The box below contains a few paragraphs. As will be shown in the pages that follow, text like this from a well-organized book can serve as the basis for numerous test items.

Supporting Information: Attenuation

Attenuation is the reduction in the number of x-ray photons remaining in the beam after the beam passes through a substance. It is the result of the x-ray beam interacting with matter and losing energy, or photons, through these interactions. There are five mechanisms by which x rays interact with matter: coherent scattering, Compton effect, photoelectric effect, pair production, and photodisintegration. Of the five types of interactions, two are important for the production of radiographs: the Compton effect and photoelectric absorption.

The Compton effect, also called Compton scatter, occurs when an x-ray photon interacts with an outer-shell electron in body tissue, which causes the photon to lose energy and change direction, or scatter. The outer-shell electron is ejected from the atom. The ejected electron is called a secondary electron or Compton electron. Some of the scattered radiation may reach the film. Because it is of low energy and strikes the film from different directions, it decreases radiographic contrast by creating a uniform darkness on the film, called fog. Although Compton scatter is typically low energy, it is still sufficient to require that x-ray examining rooms be shielded.

Photoelectric absorption results when an x-ray photon interacts with an inner-shell electron. The photon is not scattered but is absorbed by the substance, resulting in a reduction in the number of photons that pass through the substance. The inner-shell electron that is released from the atom is called a photoelectron.

One factor that determines the amount of attenuation is the type of irradiated substance. Thicker body tissue results in greater attenuation, as does more dense body tissue. When density is doubled, the chance of x-ray interaction is doubled because twice as many electrons are available for interaction, and more photons will be absorbed. Attenuation is also affected by the atomic number of the substance because substances with high atomic numbers have more electrons available for x-ray interaction. Three cm of bone will attenuate more of the x-ray beam than 2 cm of bone, and 3 cm of bone will attenuate more than 3 cm of muscle.

X-ray photons that penetrate the body and reach the radiographic film produce dark areas, called areas of high optical density. The photons that are attenuated by the body and do not expose the radiographic film result in light areas, or areas of low optical density. The process of attenuation is what causes differences in tissue composition to show up as differences in optical density on an x ray.

2. Determine the candidate's task

Once a topic has been identified, your next step is to consider what you expect the candidate to do with it. Given that written exams are limited to testing for cognitive knowledge and skill, the candidate's task is a mental, not physical, activity. So, this step is really about deciding whether the candidate should have to remember some fact, apply a formula, interpret patient data, evaluate a radiograph, and so on.

Begin by clarifying the topic. In the box on the preceding page, attenuation is the general topic, but it comprises several specific pieces of knowledge — and it's these specific units of knowledge that most test items are really about. These smaller units of knowledge can be *facts, concepts,* or *principles*.

Facts are simply statements taken to be true. Like these:

- Attenuation is the reduction in the number of x-ray photons remaining in the beam after it passes through a substance.
- The ejected electron is called a secondary electron.
- Fog is unwanted density uniformly distributed over the film.

Concepts are characterized by their defining attributes. For example, the concept *bird* is defined by attributes such as feathers, two legs, wings, egg laying, and so on. The concept of *radiographic quality* is defined by attributes such as density, contrast, recorded detail, and so on. Concepts may also have examples: a radiograph with static marks might be presented as an example of a radiograph lacking in radiographic quality due to unwanted densities.

Principles involve cause–effect relationships and often take the form of "if … then" statements. Radiologic technology is driven by principle. The inverse-square law is one very well-known principle; the relationship between kVp and optical density is another. Among the principles in the text box is: "*If* a substance has a high atomic number, *then* more electrons are available for x-ray photons to interact with, and more photons will be absorbed."

Once the facts, concepts, and principles have been identified, it's relatively easy to specify the candidate's task. The box on the next page presents several task statements related to attenuation. Each identifies — most at the beginning — the type of cognitive skill we expect to be demonstrated.

TIP:
Make sure that the task includes a verb that describes the cognitive process the candidate is expected to go through.

Sample Tasks

The candidate will:

- Recall that fog is defined as an unwanted density that is uniformly distributed over the film
- List the five types of interactions between x rays and matter
- Recall that the photoelectric effect and Compton effect have important implications for image production
- Recognize that coherent scattering typically has a negligible effect on r adiographic quality
- Recognize the definition of a secondary electron
- Explain the event that occurs when an x-ray photon interacts with an outer-shell electron
- Given a simple line drawing of an atom (nucleus, protons, electrons), indicate the electron shell at which the photoelectric effect is likely to occur
- Identify tissue density as one of the factors that affect the amount of x-ray beam that is attenuated
- Differentiate, given a line drawing of an atom, a photoelectron from a Compton electron
- Recognize why substances with higher atomic numbers absorb more photons
- Given a list of common substances (e.g., bone, muscle, water, barium), estimate which has the greatest and least amounts of absorption
- Explain how scatter results in fog
- Recall that x-ray photons reaching the film produce areas of high optical density

We could easily develop more. The important thing to recognize is that it's not necessary to actually write down all of these tasks when creating test items. But it is important to locate a good reference, read the relevant sections, and give some thought to the cognitive behaviors you seek to evaluate.

Here are a few cognitive verbs to use in your own work:
Analyze, Choose, Compare, Comprehend, Conclude, Define, Detect, Differentiate, Distinguish, Estimate, Explain, Identify, Indicate, List, Match, Plan, Predict, Recall, Recognize, Restate, Select, State, Understand.

3. Formulate the stem and correct answer

After you've considered the candidate's task, it's time to do some writing. The stem and correct answer are pretty straightforward. As we shall later see, writing distractors is the hard part!

Candidates must be given a clearly stated and complete problem to solve. Occasionally, it's even OK to include information that might lead the candidate to an incorrect answer. Although we don't want to trick them, it's perfectly acceptable to determine whether they can differentiate relevant from irrelevant information.

TIP:
The stem should include sufficient information to allow the candidate to determine the correct answer.

It's almost always best to phrase the stem as a direct question, which helps ensure that it poses a focused problem. If, after writing the stem and options, it's apparent that the stem would read better as an incomplete statement, then by all means feel free to go back and revise the stem. The best item writers start with a strategy for focusing the problem. Chapter 3 offered several guidelines related to writing the stem; follow those and you'll be in good shape.

4. Write the Distractors

The primary difficulty with writing distractors is that most of our thinking is oriented to the correct answer. When we identify information to support a topic, as in the box on the preceding page, we typically focus on correct, not incorrect or misleading, information.

Here are a few tips:

- Try to anticipate incorrect responses from uninformed and misinformed candidates. The uninformed just don't know, and any distractor works for them. (Fortunately, there aren't too many of them out there.) The misinformed know only a little about a topic, or they're confused by a misconception. This is the person who mistakes contrast for density, or forgets to invert when thinking about the inverse square law. Try to put yourself in their shoes. What are some common misconceptions about this topic? What parts are most difficult to master or easiest to forget?
- Develop distractors from irrelevant attributes — those features that define a related concept. For example, an irrelevant attribute of the photoelectric effect is its interaction with an outer-shell electron (remember that outer-shell electrons are an attribute of the Compton effect). So an item about photoelectric effect might have *outer-shell electron* as a distractor.
- If a test item calls for an example of a concept, bad examples can be transformed into distractors. Consider an item that asks for examples of the five types of interactions between x-ray photons and matter: it might use distractors like *fusion, induction,* and *differential absorption.*
- It's usually pretty easy to write distractors for principles, because they deal with relationships, usually "if…then" relationships. An item that asks about the effect of kVp on radiographic contrast could have options such as *increases, decreases,* or *stays the same.* These options all deal with the direction of the relationship. Alternatively, options can be about the strength of the relationship (e.g., "Doubling X will have what effect on Y? — halve, double, quadruple"). Options may also pertain to the shape of the relationship (proportional, geometric, logarithmic, inverse, and so on). For certain types of questions, it's impossible to come up with four reasonable options. In some instances, it's OK to have three, such as *increases, decreases,* and *stays the same.*
- Another feature of principles is that they are often expressed mathematically. Recall that the inverse-square law states that the intensity of radiation at a given distance from the source is inversely proportional to the square of the distance, or, $I_2 = I_1 \times D_1^2 / D_2^2$.

An exposure of 200 mR is obtained at a distance of 20 inches. What will the exposure be if the distance is increased to 40 inches?

A. 16 mR
B. 50 mR *
C. 400 mR
D. 800 mR

Now consider this item:

The correct answer is: $200 \times 20^2 / 40^2 = 50$ mR. All of the options should come from misapplications of the principle, such as forgetting to square one or both values, taking the square root, dividing instead of multiplying, and multiplying instead of dividing. For this example:

- distractor A was obtained by: $40^2 \times 20^2 / 200^2 = 16$
- distractor C was obtained by: $200 \times 40 / 20 = 400$
- distractor D was obtained by: $200 \times 40^2 / 20^2 = 800$

- For some math items, plausible distractors can be created by moving the decimal, using parentheses incorrectly, or varying the number of zeros. Consider these types of distractors if the item involves converting from one unit of measure to another (e.g., SI units to conventional units), when dealing with the metric system, or for items that require scientific notation.

The bottom line: for distractors to appear plausible, they must be thoughtfully crafted. Don't just make up wrong answers; derive them by attempting to mimic common misconceptions and mistakes. Several sample test questions that illustrate these points appear in the following pages.

5. Review the Item

Just as important reports or letters require careful proofing and editing, so do test items. And the benefit is twofold: first, it improves the item; second, the revision process can generate even more items on the same topic. Item review should be guided by the checklist presented at the end of Chapter III.

Putting It All Together

Remember the paragraph on attenuation and the following student task statements? Here's where we convert them into test items. Several items are presented in the box; along with comments.

This item evaluates the student's ability to recognize some fundamental facts about types of interactions relevant to radiography. Although pretty basic, it's a good solid item. Notice that the distractors all correspond to related concepts taken from other task statements.

Which of the following are considered to be the two most important types of interactions for radiography?

A. coherent scattering and photodisintegration
B. pair production and photodisintegration
C. Compton effect and pair production
D. Compton effect and photoelectric effect *

This one goes beyond basic scientific facts by requiring candidates to understand the impact of certain types of interactions on an actual radiograph. This isn't necessarily a better item, but its relevance to practice is more obvious.

Which of the following types of interactions between x rays and human tissue is associated with fog on an exposed radiograph?

A. coherent scattering
B. Compton effect *
C. pair production
D. photodisintegration

Here we get away from facts and into principles, with one of those important "if.... then" relationships. Note two things about the options: first, they form pairs of opposites (generally a good approach); second, the concepts in each of the options are based on ideas taken from the list of task statements presented on page 38.

Compared to substances with lower atomic numbers, those with higher numbers:

A. absorb more energy from the x-ray beam *
B. absorb less energy from the x-ray beam
C. are associated with thicker body tissue
D. are associated with thinner body tissue

Note the distractors on this item: option A is plausible to a misinformed or confused candidate because it's the inverse of the correct answer; option D might be an attractive guess for an uninformed candidate who's heard of the inverse-square law. Admittedly, option B probably is not very effective; it may be better to play off the inverse square law (e.g., *decrease by the inverse of the square*), although the wording seems to get complicated.

As the density of a material is doubled, the probability of an x-ray interaction:

A. is reduced by half
B. remains the same
C. is doubled *
D. is squared

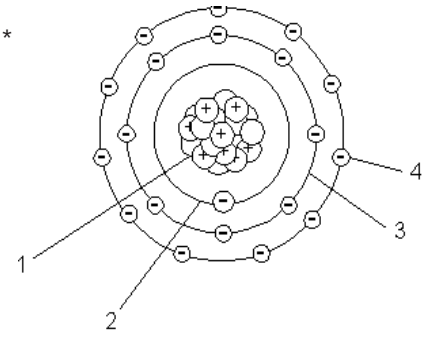
Why does dense material result in greater attenuation?

A. The electrons are larger in dense material, which blocks more of the x-ray beam
B. More electrons are available for interaction in dense material, which increases the number of photons absorbed from the x-ray beam *
C. As density increases, the size of the nucleus increases, and the larger nucleus absorbs more photons from the x-ray beam
D. The binding energy of the k-shell in dense material is higher, and this causes more photons to bind with the k-shell electrons

Here we have the same content, but more difficult — requiring the candidate to explain *why* something happens. Asking "why" is one of the advanced strategies discussed in the next chapter. The downside of this strategy is apparent in the item: asking "why" often requires lengthy options. That's OK if the item gets at important content.

Which number in the figure corresponds to the K shell?

A. 1
B. 2 *
C. 3
D. 4



Which number in the figure indicates the likely location for photoelectric absorption?

A. 1
B. 2 *
C. 3
D. 4

The first of two items in the box at left refers to a common drawing depicting interactions between x rays and matter. Simple but effective, it requires the candidate to identify a correct label for a part of the drawing. In the health sciences, items like this are slam-dunks (anatomy, positioning, drawings of equipment, and so on).

Similar to the first item, the second one shows how drawings can do more than identify parts: candidates must demonstrate understanding of a process. You can also write items about how parts interact with other parts, what happens when they malfunction, and so on (see chapter 2).

Chapter Summary

Before leaving this chapter, let's restate the five steps of item production:

1. Identify a topic and supporting information
2. Determine the candidate's task
3. Formulate the stem and correct answer
4. Write the distractors
5. Review the item

We'd also like to re-emphasize that it's really not necessary to follow this process in lock-step fashion. If you skip a step, that's fine. Merge steps together — that's OK, too. Maybe two big steps will work: 1 and 2, then 3 through 5. Make them work for you.

The important point of the exercise is this: obtain supporting information and think clearly about what you expect the candidate to be able to do with that information. This preliminary task isn't something you do for each item; rather, you do it just once for a whole set of items.

Think about it: from the handful of task statements presented earlier in the chapter, it wasn't a big chore to develop multiple items on related topics. We could have easily written another five or ten items with minimal additional effort. How did we manage that? Having all the related information summarized in one place made it relatively painless for us to finally write those downright nasty distractors.

V) Critical Thinking and Clinical Problem-Solving Skills

Chapter IV identified basic strategies for writing test items. Now we put them to work. We begin by discussing levels of cognitive complexity. Then we demonstrate strategies for producing items that assess those higher-order thinking skills. The final part of the chapter discusses test items that assess clinical decision-making skills — the types of skills that technologists exercise in their day-to-day work.

Critical Thinking

Levels of Cognitive Complexity

Many theories describe the cognitive processes used in clinical and didactic settings with terms like *prioritizing, inferring, analyzing, problem solving,* and *evaluating*. Phrases like *higher-order thinking skills (HOTS), critical thinking,* and *clinical decision making* have enjoyed popularity in recent years.

Bloom's Taxonomy — around since the 1950s — is the most widely recognized framework for describing levels of cognitive complexity. It identifies six levels of cognitive skills: knowledge, comprehension, application of knowledge, analysis, synthesis, and evaluation. This framework has been very useful, and all six levels come into play when writing test items. But we believe a simpler framework, consisting of just three levels of cognitive complexity, works just as well:

1. Recall/Recognition. Test items at this level require candidates to recall or recognize previously learned facts, concepts, and principles. Examples of a cognitive skill at this level: listing the bones of the foot on an anatomical drawing; being able to restate the inverse-square law verbatim.

2. Application. Test items of this nature require candidates to apply previously learned information to a practice-related problem. Calculating new exposure factors based on application of the inverse-square law is an example of this type of test item.

3. Problem Solving. Items at this level involve analyzing situations, evaluating information, and determining solutions to problems. Candidates are required to use information they already know, evaluate information that is supplied by the test item, then integrate the two to solve the problem posed by the test item. Problem solving usually requires critical thinking.

Two Important Caveats

First, these three levels aren't carved in stone, nor are they any better than the six levels proposed in Bloom's taxonomy. Whether there are two, three, or eight levels probably doesn't matter. What's important is for item writers to think about the cognitive demands of the test questions they write.

Second, a test item's cognitive complexity depends on more than just the item itself; it also depends on the candidate who responds to that question. An item that's problem solving for a marginal candidate might be simple recall for one who's well prepared. Maybe this is why studies have shown that test items classified as *analysis* or *problem solving* are no more difficult than those classified as *recall* or *comprehension*.

We thought it helpful to discuss cognitive complexity because there really is such a thing as depth of knowledge: candidates do learn different things to different degrees, and test items do vary in terms of the mental demands they place on candidates. Following are a few strategies to help you write test items that will require candidates to put on their thinking caps.

The Socratic Dialogue

It's probably safe to say that Socrates was no fan of the MCQ format. History tells us that he preferred oral exams as a method for getting at a student's true depth of understanding. That method of questioning has come to be known as the Socratic Dialogue.

The two formats are not mutually exclusive. You can write MCQs based on the Socratic Dialogue. The key is to imagine yourself sitting across the table from one of your better students.

Let's say that you've just asked a basic question, like, "Which type of tissue attenuates an x-ray beam the greatest amount?" The candidate gives the correct answer. What next? What path do you take to complete your assessment? You could ask more factual questions about other tissues or other types of radiation — this would be a reasonable approach to get at the student's breadth of knowledge. Alternatively, you could use a few strategies to get at depth of knowledge. Three strategies can turn basic recognition items into items that assess critical thinking. The strategies are: asking "why," asking "how," and asking "what if."

• ***Asking 'Why'***

The quickest way to assess knowledge at a deeper level of understanding is to require candidates to explain the rationale underlying some concept or principle. A candidate may know that bone attenuates the x-ray beam more than muscle does, but do they really know why? Alternative questions could ask the candidate to explain why some tissues absorb more than others, or to describe the conditions that result in higher levels of scattered radiation.

The trick is to ask "why" within the confines of the MCQ format. Consider the two items at right. The first simply requires students to recognize that smoking is a risk factor for osteoporosis. The second one requires some understanding of why smoking is a risk factor.

Which of the following is likely to decrease bone mineral density in premenopausal women?

A. cigarette smoking *
B. membership in Weight Watchers
C. nulliparity
D. use of birth control pills

Why does cigarette smoking result in decreased bone density in premenopausal women?

A. the decrease in oxygen supply associated with smoking decreases the oxygen available for bone formation
B. smoking over-stimulates the production of osteoblasts
C. smoking suppresses overall metabolism thereby decreasing the regularity of the bone formation cycle.
D. smoking chemically alters estrogen so that it no longer contributes to bone formation *

The second of these is a cognitively complex and difficult item. However, it illustrates a couple of potential limitations to asking why. First, such questions often require lengthy options, because each must offer a plausible explanation, and explanations usually require many words. Long options are OK; it just takes extra care to assure that they're not too wordy.

Second, asking "why" can make very hard questions out of easy subject matter. When the items above were pretested as experimental questions, we found that over 90% of bone densitometry candidates knew that smoking is a risk factor for osteoporosis, but only 30% understood why. Is *why* important in this situation? As it turns out, the Bone Densitometry Examination Committee felt that knowing *why* for this particular topic was not essential to effective job performance, and deleted the latter of the two questions from the item pool.

Which of the following substances attenuates an x-ray beam the greatest amount?

A. muscle
B. fat
C. air
D. bone *

Assume a patient has osteoporosis. How will this condition affect the interactions of x-ray with bone?

A. osteoporotic bone attenuates more photons
B. osteoporotic bone attenuates fewer photons *
C. osteoporotic bone produces more scattered radiation
D. osteoporotic bone decreases contrast

Compared to normal breast tissue, the radiographic density of an irradiated breast is:

A. radiolucent
B. more dense
C. less dense *
D. of equal density

- ***Asking 'What If'***

Nothing is completely true all of the time — sometimes the correct answer depends on certain conditions. Drawing on these conditions can open the door to additional test items. The first item in the box at left is a simple one about the attenuating effects of different types of tissue. This topic can be made more complex by asking about the effects of different types of bone or of various pathologies. For example, what if the patient has received radiation therapy, or is postmenopausal? How would these factors influence density and attenuation? The latter two questions arguably require a deeper level of understanding than the first.

How might the radiographic technique be modified for an AP hip for a patient diagnosed with severe osteoporosis?

A. increase exposure factors
B. decrease exposure factors *
C. no change to exposure factors

What changes in exposure factors may be required for elderly patients scheduled for a mammogram?

A. increase kVp
B. decrease kVp *
C. increase focal spot
D. decrease focal spot

- ***Asking 'How'***

You can determine if candidates know how some piece of information applies to their jobs. A related strategy is the *so what* line of questioning, which requires candidates to understand how certain things impact practice. The items at left assess the ability to adapt and apply knowledge to nonroutine situations.

So far, this chapter has addressed primarily didactic knowledge. Now, we turn to the assessment of clinical skills. Although the strategies just presented (asking why, how, and what if) also apply to clinical skills, a few additional tricks can be helpful when writing items that require candidates to apply their knowledge to practice-related problems.

Assessing Clinical Skills

Let's clarify what we mean by clinical skills. Listed below are a few examples of clinical activities that occur in the practice setting.

- Transport a patient with a fractured hip
- Set-up equipment for a PA chest radiograph for a pediatric patient
- Explain breathing instructions to patient for a PA chest radiograph
- Perform a bronchogram
- Perform a QC test for a collimator light

Each of these activities involves patients, equipment, or both. Each also requires some sort of psychomotor skill, and may even involve interpersonal or communication skills. Most importantly, none of these activities can be directly assessed with MCQs. Not even the most cleverly written test item can test a technologist's ability to transport a patient or set up equipment. Truly assessing such skills requires observing a technologist's interaction with an actual patient in a clinical setting with real equipment.

Impossible? No. Impractical? You bet.

The alternative is to develop clinically relevant test items. Key to writing good practice-based items is the fact that all clinical activities require some type of knowledge. In other words, you can assess clinical skills by assessing the knowledge and cognitive skills that underlie the procedures. Here are four sources:

1. Steps of a Procedure

MCQs can determine whether a candidate can identify the appropriate steps and place those steps in proper sequence. Alternatively, candidates can be given the steps and asked to name the procedure or explain its purpose. The item at right requires candidates to know that anesthetic is given prior to doing a bronchogram. The next two items pertain to QC tests: one for a compression device on a mammography unit, and the other for a collimator light on a radiographic unit.

When performing a bronchogram using an intratracheal catheter, what is usually done prior to inserting the catheter?

A. a local anesthetic such as lidocaine is administered to the larynx *
B. a tracheotomy is performed and a temporary port is placed
C. a general anesthetic is administered so that the patient remains unconscious
D. a barium swallowing study is performed to confirm esophageal structure and function

What is the first thing to do when using a bathroom scale to perform a compression test on a mammographic unit?

A. activate and deactivate the compression device 3 times to warm it up
B. place the scale between the cassette and the top compression paddle
C. place a towel on the cassette *
D. place a phantom on the cassette

A quality control test requires the steps 1– 5 as summarized below. What type of equipment malfunction will this QC test detect?

1. place film on table of x-ray unit
2. turn on collimator light to illuminate film
3. place 4 pennies at corners of light field
4. expose film at 40" SID
5. process film and inspect

A. faulty exposure switch
B. improper SID setting
C. defective mA meter
D. inaccurate collimator light *

To correctly answer the first item in the box at left, the candidate must know not only that the towel needs to be placed on the cassette before the scale, but also what equipment is used (e.g., towels are, phantoms aren't). The *1999 ACR Mammography Quality Control Manual* clearly describes the steps of this procedure.

Answering the second item really involves critical thinking — at least for most candidates who haven't actually tested a collimator light in this perfectly legitimate, but somewhat novel, way. The candidate must infer the purpose of the QC test from the steps that are given. The downside to this item is that it consumes a lot of space and requires a lot of reading. But such items are worth the effort if written well.

2. Concepts and Principles Underlying a Procedure

Although test items about the steps of a procedure are useful, they often require little more than recall of facts. It's possible to move beyond recall, with items that assess:

- How a certain step is performed
- Why a certain step is performed; why it is needed
- Why one step occurs before or after another
- Types of equipment and instrumentation required; knowledge of that equipment (e.g., how it works)
- Anatomy, physiology, or pathology involved in the procedure
- Critical things to be careful of when performing a step
- What happens if a certain step is omitted, performed incorrectly, or performed with inadequate instrumentation
- Indications or contraindications for a procedure

The items at right assess some of these cognitive skills. The first two pertain to an ERCP, while the remaining three are based on the mammography compression test. Most require the candidate to demonstrate fairly detailed knowledge about the steps of a procedure by asking *why, how,* or *what if.*

When performing an ERCP, contrast is used to enhance the biliary ducts. To introduce the contrast, it is necessary to first locate what anatomical structure with the endoscope?

A. common bile duct
B. ampulla of Vater
C. islets of langerhorn
D. sphincter of Oddi *

When performing an ERCP, why is the larynx anesthetized prior to inserting the endoscope?

A. to prevent esophageal reflux when contrast is introduced
B. to minimize superimposition of the thyroid cartilage
C. to minimize interference from the tongue
D. to prevent gagging when the endoscope is introduced *

Why is it necessary to place a towel on the cassette holder before performing a compression test on a mammographic unit using a bathroom-type scale?

A. to prevent damage to the cassette holder *
B. to protect the scale
C. to simulate the compressibility of breast tissue
D. to calibrate the unit before performing the compression test

A mammography unit should be subjected to a compression test whenever inadequate compression is suspected, and every:

A. day
B. week
C. month
D. 6 months *

What type of scale is effective for a compression test on a mammographic unit?

A. digital
B. analog *
C. weighted counterbalance, such as a chemist's scale
D. digital scale for filmless mammography units; analog scale for film mammography units

Why should an analog scale be used instead of a digital scale when performing a compression test on a mammographic unit?

A. analog scales require calibration less often
B. digital scales sample the data and may not respond accurately as compression is slowly applied *
C. analog scales are more accurate in the compression range of interest, which is 30 to 50 lbs
D. digital scales are too expensive to be practical

A compression test on a mammographic unit consistently gives readings of 21, 22, and 23 lbs on three testings obtained within a few minutes. What should be done?

A. record the average value, and repeat the test at the next scheduled interval because the values are within acceptable limits
B. record the high value, and repeat the test at the next scheduled interval because the values are within acceptable limits
C. contact a service engineer or physicist because the values are not within acceptable limits *
D. recognize that the measurements are inconsistent, and repeat the test on the next working day to confirm the results

The PSA level for a 63-year-old retired male is found to be 3.1. His general health appears to be normal and nothing remarkable was noted on the digital rectal exam. What would be the next step in patient care?

A. Have the patient return in a year for follow-up *
B. Perform another digital rectal exam
C. Order a transrectal ultrasound (TRUS) of the prostate
D. Perform a needle prostate biopsy

3. Clinical Data or Procedure Outcomes

Procedures result in a product that's either an end in itself or used as input to some other procedure. Here are a few of the related cognitive skills that can be assessed with standard MCQs.

- Interpreting the results of, or output from, a procedure
- Evaluating results or output for quality (e.g., radiographic contrast, distortion, etc.)
- Recognizing limitations of the process or the instrumentation involved
- Knowing what and how to document

Three items at left illustrate these ideas. The first two pick up on the mammography QC theme. One requires the candidate to differentiate between digital and analog scales for measuring compression and to recognize the limitation of digital scales. The other requires candidates to evaluate the outcome of the QC test and justify a course of action. MQSA requires it; ACR requires it; so they better know it!

The third item, from the field of radiation oncology, requires the candidate to interpret clinical data, differentiate normal from abnormal values, and determine a course of action. This item is probably getting into the domain of the medical practice, but nonetheless illustrates the type of clinical problem solving that MCQs can test. For example, an item might require evaluating radiographic quality by presenting a mammogram and asking the candidate to decide what view to do next.

4. Unexpected Circumstances: What to do Next

Here's a twist on a strategy that was briefly mentioned earlier. Sometimes, for various reasons, things go wrong. It could be that the patient does something unexpected, has variant anatomy, or is limited by injury or pathology. Equipment problems may also require a technologist to respond quickly.

The two items at right ask candidates what to do next. Keep in mind that items like this often require common sense or sound clinical judgment, and it can be challenging to write distractors that are plausible but incorrect. Another challenge is that opinions can vary: what's viewed as appropriate by one expert might be considered a waste of time by another. Even so, this type of item is often worth the effort.

A patient for a routine chest radiograph exhibits seizure-type behavior, and then falls to the floor. He appears to be unconscious, but breathing. What should be done first?

A. check for head injuries due to the fall
B. obtain x rays for any regions where injury is suspected
C. attempt to arouse the patient
D. call or send for a physician as quickly as possible *

The PSA level for a 63-year old retiree is found to be 19.7 ng/ml on a routine screening. The digital rectal exam was mildly suggestive of enlargement. What would be the next step in patient care?

A. Have the patient return in a year for follow-up unless he notices significant changes in GU function before then
B. Perform another digital rectal exam
C. Order a transrectal ultrasound (TRUS) of the prostate
D. Perform a needle prostate biopsy *

Summary

This chapter presented a simplified cognitive taxonomy consisting of three levels of processing: recall/recognition, application, and problem solving. Although it seems that most items are written at the recall level, it's not difficult to write items that involve application or problem solving. Strategies for writing items at the higher two levels include asking "why," "what if," and "how."

We also discussed the nature of clinical problem-solving and how MCQs can be written to get at clinical skills. Although it's not feasible to write items that directly evaluate a candidate's ability to actually carry out a procedure, it *is* possible to assess cognitive processes — like knowledge of the steps of a procedure and of the principles underlying each step; ability to evaluate the data, results, or other outcomes from a procedure; and skill at managing unexpected circumstances.

Applying the techniques covered here, many item writers find they can make their multiple-choice exams more challenging, more interesting, and more relevant to clinical practice.

The American Registry of Radiologic Technologists®

Promoting high standards of patient care by recognizing qualified individuals in diagnostic medical imaging, interventional procedures and radiation therapy

1255 Northland Drive
St. Paul, Minnesota 55120
Telephone (651) 687-0048

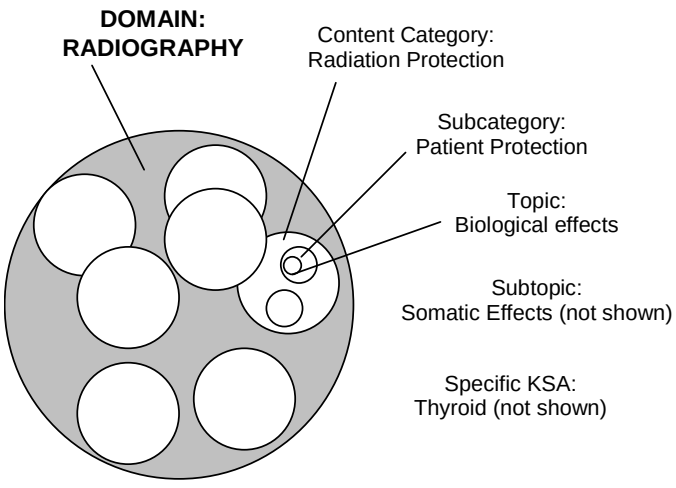
Visit our web site at www.arrt.org

Appendix A: **The Length of Certification Exams**

This article was originally featured in ARRT's 2001 Annual Report to Registered Technologists.

An anxious candidate will occasionally call and wonder why his or her ARRT examination has *so many* questions. And then, before you know it, a genuinely concerned educator calls to ask why the exams have *so few* questions. Both questions are legitimate, and both can be answered with the same explanation.

Certification exams are often developed according to the domain sampling model. A domain can be thought of as a set of skills. Domains can be large and cover a multitude of loosely related skills or be small and include a few tightly related skills. For example, skill in internal medicine would represent an extensive domain, while knowledge of cardiac anatomy seems to be a fairly compact domain.



The Domain . . . What?

The domain sampling model recognizes that no single exam can reasonably include questions about every possible knowledge, skill, and ability (KSA) in a particular domain. Instead, any one exam is comprised of a sample of the domain. So, rather than assembling exams that consist of hundreds or even thousands of test questions, the ARRT assembles exams that typically range from 100 to 200 questions sampled from a larger domain.

The domain sampling model works only if sufficient attention is given to defining the domain and if the domain is adequately sampled. A clearly defined domain must be well conceived and carefully mapped out. This is accomplished through a practice analysis. The goal of a practice analysis is twofold: (1) to identify the tasks required for effective job performance and (2) to determine the KSAs required to perform those tasks. The KSAs make up the domain.

Once the domain is defined, a strategy for sampling from the domain must be implemented. The figure illustrates the relationships among the different levels of a domain. The more specificity built into the domain, the greater level of control one can have when sampling the domain. One important feature of a sample is its size; an exam needs to be long enough in order to reliably sample the domain. Another important feature is representation; a representative sample of test questions leads to a balanced exam.

(Continued on next page)

At the ARRT, the length of an exam is determined primarily by the Practice Analysis Advisory Committee. The Advisory Committee considers several factors including the breadth of the domain, the number of content categories, the importance of the tasks within a content category, the relatedness of content categories, and other data obtained from the practice analysis. In addition, ARRT staff draws on common psychometric practices.

We know from experience that a 500-question exam is overkill for all but the broadest of domains and that a 50-question exam is too short for all but the narrowest domains. Although there are statistical procedures for verifying the adequacy of exam length after an exam has been administered, the process of initially establishing the length of an exam is primarily a judgmental one.

The next activity is to ensure that the domain is well represented. The results of this activity are reflected in the content specifications for an exam. The content specifications indicate, among other things, the number of questions for each of the content categories, subcategories, and topics. When establishing weights for a content category, the Advisory Committee considers the information from a variety of sources including curriculum guidelines, recommendations from the professional community, and the results of the practice analysis. The number of questions allocated to a content category is strongly influenced by the breadth of the category as well as its importance to practice.

The domain sampling model generally works well for the same reasons that opinion polling works — a carefully specified sample really can tell you quite a bit about the entire population. Another fact that works in favor of the domain sampling is that test-takers' behavior is reasonably consistent from one topic to the next.

If, for example, a candidate performs well on the radiation protection section of the Radiography exam, chances are that he or she will also do well on image production and evaluation. Scores on those two sections of the Radiography exam are highly correlated.

Meanwhile, scores for two seemingly unrelated content categories (such as image production & evaluation and patient care) still exhibit a moderate correlation. This consistency in performance on different parts of a domain bodes well for domain sampling.

Yeah, Right ... But Does it Really Work?

When a person takes a test and earns a particular score, we want to be confident that the score is a reliable indication of that person's knowledge for the entire domain. Let's say Joe gets a score of 80% on a 25-question radiography exam one week. A week goes by and Joe takes another exam covering the same domain.

Unless Joe did some serious studying or was for some reason not himself the week before, we would expect him to get a score near 80% on the second exam. If Joe scored 60% on a second exam and 90% on a third we would become suspicious because the scores don't seem reliable.

At this point, we could blame Joe or we could blame the exam. If there were a lot of Joes taking a lot of exams, and their scores bounced around by more than a few points, then we should get really suspicious of the exam. Although many factors could contribute to score instability, an incomplete sampling of the domain is one of the more obvious ones. If you don't ask enough questions, you're not going to get reliable scores.

There are several ways to evaluate score reliability. The most common methods are based on the extent to which scores on tests from the same domain correlate with one another. Calculating reliability is like giving many tests from the same domain to many Joes. Reliability indices can fall between 0 and 1, and an index of .90 can be considered as a good target for certification exams.

To evaluate the impact of test length on score reliability, we can do a short study of the ARRT Examination in Radiography. The data for this particular investigation were obtained from the October 1998 exam, which was taken by 2564 first-time candidates. The Examination in Radiography consists of 200 questions. The total test reliability index for this group was 0.926.

For this study we will see what happens when exam length is decreased by artificially making the exam shorter. The study involves four steps: (1) randomly discard 20 questions from the exam; (2) recalculate the reliability index to see how much it goes down; (3) see if the pass/fail decisions for any candidates have changed from fail to pass or from pass to fail; (4) repeat the process for successively shorter exams by discarding 40, 60, 80, and 100 items.

Show Me The Money

The table below shows the results of our investigation. It looks like tossing out 20 to 40 questions does not have a dramatic impact on test reliability. It is not until after exam length falls below 140 that notable changes in reliability start to occur.

Number of Questions	Reliability Index	Decision Changes P to F	F to P
200	0.926		
180	0.920	0.3%	0.7&
160	0.914	0.5%	0.9%
140	0.903	0.8%	1.3%
120	0.888	1.1%	1.3%
100	0.864	1.1%	2.3%

However, the reliability index does not tell the complete story, particularly for individual candidates. The table indicates that pass/fail decisions for some candidates would have changed if a shorter radiography exam had been given. An exam consisting of 180 items would change the decisions for 1% of the group (0.3% + 0.7% = 1.0%), while an exam consisting of 100 items would affect over 3% of the group.

The changes in pass/fail decisions can be regarded as the errors that would have resulted from shortening the radiography exam. Nonetheless, it is interesting that the exam could be cut in half (to 100 questions) and still produce the same pass/fail decision for almost 97% of the candidates.

In general a longer test is a better test. However, for most tests, there is a point of rapidly diminishing returns after exam length reaches about 150 items. The advantages of adding questions to an exam needs to be weighed against costs such as test development expenses and testing time, both of which are related to examination fees. Candidate fatigue also becomes a factor for very lengthy exams. Finally, longer exams means that more questions are exposed to test-takers, and for Boards like the ARRT that reuse test questions, overexposure of questions is a security concern.

But What If . . .?

This previous discussion may have prompted more questions than it answered, so let's address a few of them now.

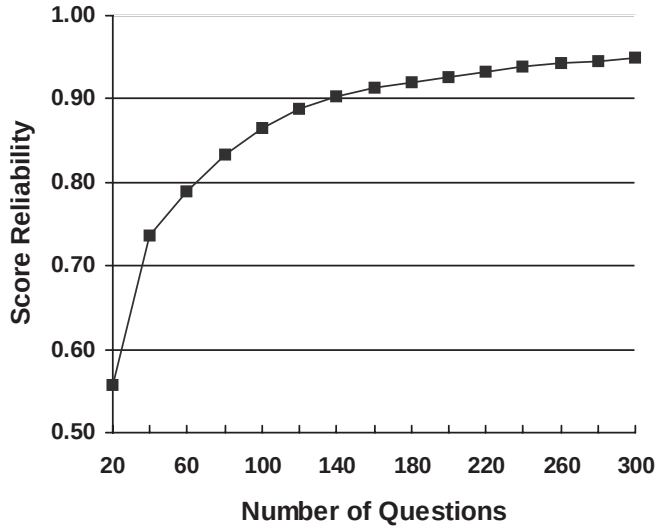
Question: What would happen if the Radiography exam was cut to 20 questions or increased to 300 questions?

Answer: Although the study reported in the table at left would have allowed us to drop as many questions as we chose, there was no way that we could actually add questions. That's OK because there is a formula for estimating the impact of adding or deleting questions. The graph below shows the relationship between test length and reliability for exams. The curve illustrates an important point: the effects of adding or deleting items depends on how many items you start with. Taking 100 items away from a 150-item exam wreaks havoc on measurement precision, while adding 100 items to a 150-item exam makes only a minimal contribution.

Question: Is the optimal number of questions for an exam usually around 200? Does the point of diminishing returns occur at about 150 items?

Answer: No. The ideal exam length varies according to the breadth of the domain, the similarity of the categories within the domain, the quality of the test questions, and other factors. For more compact domains the ideal length would be much lower than 200, while for broader domains the ideal length may be greater.

(Continued on next page)



Question: The ARRT reports scores for individual sections of the exams, and yet these sections may have only 30 or 40 test questions. How reliable are these section scores?

Answer: The short answer is that section scores based on content categories are not very reliable. For example, on the Radiography exam the radiation protection section (30 questions) has a reliability of around 0.70. Although this is not too bad, it certainly is not adequate for making pass/fail decisions. This is why the ARRT uses only total scores for determining pass/fail status, and provides section scores only as a general guide to a candidate's strengths and weaknesses in specific content categories. The key is to not overinterpret small differences in section scores.

Question: Just why are some exams so long?

Answer: Sometimes for good reason, especially if the domain is very broad. However, exams for some professions are unnecessarily long. One reason is that a longer exam may appear to the casual observer be more valid than shorter exams. Certain traditions seem to endure — even when such traditions may be contradicted by empirical evidence. The bottom line is that although longer exams can provide better content coverage and score reliability, there is a point of rapidly diminishing returns. And it is important to strike a balance between the costs and benefits of greater length.

Appendix B:
Item Review and Critique

These sample items identify problems and provide solutions.

The amount of radiation necessary to produce a noticeable skin reaction is called an:
A. depth dose
B. erythema dose*
C. filtered radiation dose
D. irradiation dose

Problem: Since the stem ends in *an*, the correct answer must begin with a vowel.

Solution: Some might change the stem to *a(an)*, but we suggest adding *a* or *an* to each option.

What is the normal kVp range used in mammography?
A. 20 - 25 kVp
B. 30 - 35 kVp *
C. 25 - 40 kVp
D. 35 - 50 kVp

Problem: the options overlap; thus *C* is partially correct.

Solution: Ranges should not overlap. Also, try to make them the same width, like: 30-35; 40-45. (Note: Overlap is not always a problem. Example: When the correct answer is an exact range, and only one option has that range.)

What is the primary reason for spreading a treatment dose over a number of applications rather than giving it all at once?
A. normal cells repair themselves more quickly than cancer cells *
B. it makes less efficient use of staff time since multiple visits are required
C. evens out the work load
D. multiple exposures increase the probability of negative side effects

Problem: Options B and D are not plausible. The stem asks for an advantage, but these options are obviously disadvantages.

Even option C is a bit silly. Work load should never be the reason for choosing a method of treatment.

In sum, anyone with common sense would choose A. So, the three distractors should be rewritten.

Artifacts that occur in ultrasound imaging include all of the following EXCEPT:
A. scanning the wrong area*
B. demonstrating posterior enhancement
C. producing acoustic shadowing
D. slice thickness

Problem: The stem is a negative; it's OK but not optimal. The real flaw is in the options. On one hand, option A has to be correct because it is not even an artifact.

On the other hand, the stem tells us to look for an artifact that does not occur, so option A is not a logical choice because it is not even an artifact. Boy, is this confusing!

Solution: Clarify the intent of the stem and rewrite option A so that it is an artifact, but one that does not occur in ultrasound.

(Continued on next page)

Three of these sentences about the prostate are correct. Which one is incorrect?

A. the prostate capsule is indistinct from the surrounding fascial tissue
B. the normal average size prostate is 4 x 3 x 3.8 cm
C. the seminal vesicles are paired structures
D. there is only one vas deferens and it originates from the epididymis *

Problem: Just another example of an unfocused stem and wandering options. One option pertains to size, while each of the others deal with other anatomical features.

Solution: Pick an option and build the stem and new options around it. Examples: *What is the approximate size of the prostate?* Then all options would pertain to size. Or, *Where do the vas deferens originate?* Here all the options would list alternative origins.

An 18-year old patient with a leg injury from an automobile accident is referred to the radiology department by the ER physician. The written request specifies radiographs to determine if the kneecap is fractured. What is the medical term for kneecap?

A. mandible
B. femur
C. patella *
D. flabella

Problem: Don't you just love it? Although all that information in the stem sounds clinically relevant, it is just window dressing. It makes the stem distracting and hard to read. This might be OK for a reading test, but not for not for a certification exam.

Solution: Drop everything from the stem except the last sentence.

What bone is juxtaposed to the radius?

A. ulna *
B. humerus
C. scaphoid
D. tibia

Problem: Why use a term like *juxtaposed*? This is another item for the reading comprehension test.

Solution: Replace *juxtaposed* with *adjacent to, next to,* or *parallel to*. Or use medically relevant language like *medial, lateral, proximal, distal.*

All of the following should be avoided by patients receiving head and neck radiation EXCEPT:

A. alcohol
B. high caloric foods *
C. spicy or acidic foods
D. dry or coarse foods

Problem: Is it necessary to state this item in the negative? It even has a double negative of sorts *(avoided… except)*. Options C and D each have two elements connected by the word *or,* which further complicates things. All in all, this item has 2 negatives, 1 *and*, and 2 *ors*.

Solution: Rewrite the stem to: *Which of the following should be recommended for a patient receiving radiation therapy to the neck?* Then carefully verify that the wording of options C and D is exact.

To increase percentage depth dose, one could:

A. decrease treatment distance
B. remove the filter
C. lower tube current
D. increase treatment distance *

Problem: Options A and D are paired opposites, which makes them attractive to testwise students.

Solution: Revise the other options so that they form a pair, like:

increase filtration, decrease filtration OR

increase treatment time, decrease treatment time.

Which of the following is commonly given for relief of minor pain:

A. an analgesic
B. heparin
C. acetaminophen *
D. Vesprin

Problem: Options overlap: C is a subset of A. If C is correct, then A also has to be correct.

The real problem is that the options are written at different levels of specificity. Option A is a drug class, B and C are generic names, and option D is a trade name.

Solution: Either stick with classes (analgesics, corticosteriods) or generic names (acetaminophen, diphenhydramine).

A technologist in your department informs you the automatic exposure control seems to be malfunctioning. You inspect the equipment and agree. What should be done next?

A. perform the necessary adjustments
B. tell the technologist to use manual exposure based on technique charts
C. notify the radiation physicist *
D. notify the department manager

Problem: Items like this are sometimes institution specific. Option C is keyed as correct, but option D might be correct for many institutions. Some facilities night even have service engineers.

Solution: Be careful about items that get into practice activities that legitimately vary from one setting to the next. Such items might be phrased like, *According to the NRC,* or *According to MQSA guidelines.*

The radiographic appearance of the breast may be affected by:

A. age
B. hormonal status *
C. intake of ascorbic acid
D. number of previous mammograms

Problem: Here is a different type of "overlapping options" problem. In this case, *age* and *hormonal status* are highly related, so option A is potentially correct.

Solution: Options should be independent. Revise option A.

What is normal adult body temperature?

A. 99.4 ºF
B. 98.6 ºF *
C. 37.2 ºC
D. 37.6 ºC

Problem: OK, this is a trivial item, but it does illustrate an important point. The options have multiple units of measurement: Fahrenheit and Celsius. This is not necessarily bad — it depends on the purpose of the item.

Solution: If the intent is to determine if students know normal body temperature, then they should be asked in the temperature units they should be expected to know. If they should know both units, then two questions might be legitimate.

However, if the purpose of the item is to determine if students can convert from one unit to the other, then the stem should be rewritten, and the distractors should be derived from misusing the conversion rule.

(Continued on next page)

Radioactive materials may be disposed of by all of the following EXCEPT:

A. by transferring it to a licensed land disposal facility
B. decayed in storage, released into general waste if below specified radiation limits
C. they may be returned to the licensed supplier
D. encased in a lead container, released into general waste *

Problem: It's negatively-worded, but that's a minor issue here. The major problem is that not all options complete the stem in a grammatically correct way. They are not parallel in structure.

Option A is OK, but C surely is not. (Neither are B and D for that matter). Sharp students will wonder how can an option be correct if it does not logically complete the stem.

Solution: Put the word *by* at the end of the stem and start each option with an *ing* verb. Or, the best cure for messy options is to state the stem as a question like this:

Which of the following methods should NOT be used for disposal of radioactive materials?

A. transferring them to a …
B. decaying them in storage and releasing
C. returning them to …
D. encasing them …

To increase the depth of penetration of a sound beam a sonographer should:

A. decrease frequency *
B. increase frequency
C. increase the velocity
D. decrease amplitude

Problem: Options A and B are a pair of opposites, which means that one of them is probably correct.

Solution: Create another pair of opposites (just be sure they are not correct). Another solution is to have no pairs. For example, just change option B to *increase intensity*.

The total radioactivity of a sample is 32mCi. After 12 days the radioactivity of the same sample is 4 mCi. Its physical half-life (in days) is:

A. 8
B. 2
C. 4 *
D. 12

Problem: The options are not in logical order.

Solution: Arrange options from lowest to highest. Also consider adding *days* to each option for ease of reading.

When dealing with an asthmatic patient the sonographer should do what?

A. place the patient in the Trendelenburg position
B. remain calm and confident *
C. continue scanning while help comes
D. prohibit the patient from taking his own medications

Problem: The stem is vague. Apparently, the asthmatic patient has a reaction, although the stem does not say this. Also, the options are a problem. Option B is a give-away. In fact, it could be the correct answer for any number of questions.

Solution: Revise the stem to something like: *A patient experiences an asthmatic episode during a routine abdominal scan. What should be done?*

Comment: It's tough to write distractors for "common sense" patient care items like this. Often, the incorrect answers are too obviously incorrect. Try to write on topics that can be backed up by scientific principles (e.g., normal values, body mechanics), best practice guidelines, or rules and regulations.

A patient for a GI study requires special dietary instructions for a scan to be performed the next day. When communicating the instructions, what should the technologist do to ensure that instructions will be understood by the patient?

A. visual contact between patient and self *
B. good posture
C. dress appropriately
D. occasional touching for emotional support

Problem: This is similar to the last item in that it requires only common sense, and has only vague support for the correct answer. Furthermore, the best answer seems to be missing (asking the patient to restate the instructions).

Another problem is that the options are worded poorly. If the item is kept, some options would be rewritten to include a verb (e.g., *maintain good posture*).

Which of the following may result from a whole-body, single dose exposure exceeding 250 roentgens of x-radiation?

1. cataract formation
2. carcinogenesis
3. genetic deformities

A. 1 & 2 only
B. 1 & 3 only
C. 2 & 3 only
D. 1, 2 & 3 *

Problem: Because option D is correct, this is essentially an "all of the above" item. These items should be used sparingly, if at all.

The combined response format is particularly troubling when the stem contains words like *may*, *might*, or *could*. A bright student who chooses A could correctly argue that there is no evidence of genetic defects caused by doses of radiation in this range. However, the instructor could counter by pointing out that the stem says *may;* and you never know, there could be some chance of a genetic deformity, no matter how small.

In short, there is an element of unfairness to this format whenever some of the options can be regarded as partially correct.

Solution: Rewrite the stem by getting rid of *may*. Consider using the standard one-best-answer format. Option 3 is controversial, so get rid of it.

Many people tend to avoid cancer patients due to their fears of:

1. cancer being contagious
2. saying the wrong thing to the patient
3. reacting negatively to physical changes in the patient

A. 1 & 2 only
B. 1 & 3 only
C. 2 & 3 only
D. 1, 2 & 3 *

Problem: Here is another item that is difficult to prove. It is based on anecdote, common sense, and informed opinion. An item writer could probably make up a list of 5 or 10 "fears" and most would be correct in some vague way.

Although it is important for students to be sensitive to the issues represented in this item, it is a poor item for a certification exam. It does not hold up to scrutiny. What does the word *many* in the stem mean? Is option 2 (*saying the wrong thing*) really true? How about option 3? What is the evidence supporting any of these options?

Solution: If items are based on opinions, it is helpful to clarify whose opinions, like: *According to the article by Smith, many people avoid cancer patients due to their fears of*:

Appendix C: **Evolution of an Item... Into a Test**

The main chapters of this book covered a lot of material related to item writing, with each chapter addressing just one piece of the entire puzzle. The purpose of this appendix is to put it all together by illustrating how one instructor developed a complete quiz on a single topic. To write the items, the instructor (we'll call her Terri) not only relied on the strategies discussed in the chapters, but also brought some of her own creativity to bear on the process.

Imagine . . .

It is 10:00 p.m., and somewhere in her office sits a dedicated clinical instructor named Terri. Although she appears to be staring blankly at her computer, Terri is deep in thought, pondering the mysteries of item writing. She is preparing to teach a unit on prostate cancer. It will be her students' first serious exposure to this topic. They will read an article,* review a chapter in a textbook, and spend time in the clinic. To assure that students have mastered the material, Terri plans to develop a 10-item quiz.

She has read the materials, but it was quite some time ago. So, before beginning the item-writing process, she digs out the assigned readings and does a thorough review. She makes a few pages of notes on essential facts, concepts, and principles. Then she recalls step two of the 5-step process: determine the student's task. So, she formulates some objectives — the cognitive tasks she expects students to master.

- Understand the anatomy of the prostate
- Restate the incidence of prostate cancer, for US population and by ethnic background
- Identify the methods commonly used to diagnose and stage prostate tumors
- Differentiate normal from abnormal PSA values
- Discuss factors that can cause PSA results to be inaccurate

At 11:30 she decides it's time to start writing. The first item is on basic anatomy, so it rolls off the keyboard fairly easily. One objective down, four to go. The second objective gets into epidemiology (more factual information), so items 2 and 3

* Many of the test items in this appendix are based on material covered in Recent Trends in Prostate Cancer and Brachytherapy Treatment, by J. Newman and A. Rolfo, published in the fall 1999 issue of *Radiation Therapist* (vol 8, no. 2, pp. 147-169).

also flow pretty easily. It's a good start. These items are fairly basic, and Terri realizes this. But she also realizes that the third objective — staging and diagnosis — is more challenging for students, so the difficult items are yet to come.

1. Where is the prostate normally located?
   A. posterior to the urogenital diaphragm and anterior to the urethra
   B. superior to the bladder and posterior to the urethra
   C. inferior to the urogenital diaphragm and anterior to the rectum
   D. inferior to the bladder and anterior to the rectum *

2. Approximately what percentage of men will develop prostate cancer in their lifetime?
   A. 5%
   B. 10% *
   C. 25%
   D. 50%

3. What ethnic group has the highest incidence of prostate cancer in the US?
   A. African American *
   B. Asian
   C. Caucasian
   D. Hispanic

Given the importance of prostate cancer, she believes that graduating students should know something about a good diagnostic work-up, even though that topic is somewhat advanced and may not be covered on the ARRT exam. She writes the four items below to address the third objective. Items 4 and 5 are fairly general questions about staging, and both would be legitimate to ask on a class quiz.

4. What method is typically used to stage prostate cancer?
   A. Gleason scale
   B. PSA testing
   C. FIGO classification
   D. TNM system *

5. The purpose of staging is to determine:
   A. the extent to which the cancer has spread *
   B. size of the tumor
   C. rate at which cancerous cells reproduce
   D. origin of the tumor

6. The Gleason scale is used to document tumor:
   A. origin
   B. grade *
   C. stage
   D. metastasis

7. What is a patient's Gleason score based on?
   A. origin of tumor
   B. tumor volume
   C. degree of tumor cell differentiation *
   D. extent of metastasis to other regions

Items 6 and 7 cover the Gleason scale. Although these two items are very similar, there is an important difference: item 6 simply requires students to know that the Gleason scale corresponds to tumor grade, while item 7 requires students to know that tumor grade is based on cell differentiation. Students who correctly answer the second of these will most certainly get the first correct. However, some students who know the first may not have a clue about the second. For now, Terri decides to keep them both on hand, but will use just one of them on the quiz.

It's now 12:02 a.m. Seven items done; that's about 5 minutes per item. Not bad. But the problem, Terri thinks, is that these items seem so academic and...and didactic. It would be nice to have some questions that get at diagnostic reasoning. This leads to the next objective: differentiating normal from abnormal PSA values. It occurs to Terri that she also expects students to know the appropriate course of action given a certain result. That thought gives rise to two items, with the second being almost identical to the first.

8. On routine screening, the PSA level for a 61-year old patient is found to be 19.7 ng/ml. The next step in patient management is to:
   A. have the patient return next year for routine follow-up
   B. perform a digital rectal exam (DRE)
   C. order radioimmunoassay for prostatic acid phosphatase (PAP) testing
   D. request that a biopsy be performed *

9. A 61-year old patient's PSA level on routine screening is found to be 19.7 ng/ml. The next step in patient management would be to:
   A. perform a digital rectal exam (DRE)
   B. perform prostate needle biopsy *
   C. recommend that the patient begin radiation therapy
   D. recommend brachytherapy for the patient

The only real difference between these items is in the options. In the first item, all options pertain to diagnostic actions. In the second, however, options A and B involve diagnostic choices, while C and D involve treatment. The second item might be important if the instructor's goal is to make sure that students don't prematurely jump into treatment prior to completing the necessary diagnostic studies. Terri plans to use the first item this year and save the second one for use next year. It also occurs to her to maybe use a much lower PSA result for next year, and have the correct answer be, "have the patient return for follow-up next year."

Terri now turns to the last objective: factors that can cause PSA results to be inaccurate. Item 10 requires the student to recognize that a biopsy can interfere with PSA testing. Although this item involves little more than recall of information, the process of writing it stimulated two additional items. Item 11 might be thought of as a critical-thinking version of number 10. Here, students must explain *why* PSA testing should be performed prior to biopsy. Note the two sets of parallel options: options A and B address sensitivity and specificity, while C and D get at physiologic response.

10. Which of the following may cause a temporary increase in PSA levels?
    A. recent prostate needle biopsy *
    B. recent digital rectal exam (DRE)
    C. chronic prostatitis
    D. moderate physical exertion

11. Why is it important to obtain blood samples for PSA testing prior to prior to performing prostate needle biopsy?
    A. PSA testing has greater sensitivity than needle biopsy
    B. PSA testing has greater specificity than needle biopsy
    C. the biopsy may suppress PSA levels
    D. the biopsy may elevate PSA levels *

12. During routine physical exam, a 46-yr old complains of increased urinary frequency over the past several months. Family history includes prostate cancer (father and father's brother); the patient has no male siblings. The patient's digital rectal exam is unremarkable. What is the next step in this patient's work up?
    A. recall the patient for another physical in one year
    B. order PSA testing at earliest possible time *
    C. perform prostate needle biopsy at earliest possible time
    D. perform prostate needle biopsy followed by PSA testing at earliest possible time

Item 12, the third one in the set, requires reasoning. Students must apply recently learned information to solve a clinical problem. This item is of the "next step" variety covered in chapter 5. It requires students: to determine that PSA testing is warranted given the patient's symptoms and family background; to recognize DRE is not always sensitive; and to understand that PSA testing should not be performed after needle biopsy.

Terri decides to keep all three items. She'll use the first one this year and see how it works. The others will be available for the final exams at the end of the semester, or to use when she teaches this unit next year.

It's 12:45 a.m. Even though Terri has covered all of the objectives she set out to, she is not quite ready to quit. Given the similarity among some of the items, they can't all be used on the same exam, so she really needs to produce a few more. Terri does not have specific objectives in mind for the additional items, but does recall reading about transrectal ultrasound in the article. She writes two more items — one involving basic recall, and the other using the "asking how" strategy to get at critical thinking.

13. Which of the following is NOT used in prostate cancer screening?
    A. transrectal ultrasound
    B. urinalysis *
    C. digital rectal exam
    D. prostate specific antigen testing

14. How is transrectal ultrasound (TRUS) useful in the diagnosis of prostate cancer?
    A. It can help differentiate benign prostatic hyperplasia from a malignancy in men with high PSA levels *
    B. It has greater predictive accuracy than PSA testing in men under 40 years of age
    C. it is effective in detecting small volume cancers
    D. It can determine the extent of metastases to the lower GI tract

Looking back, Terri realizes that her pool of items does not address the *treatment* of prostate cancer. Since she views this as an oversight, she puts her fingers to the keyboard one last time. She manages to tap out two more items addressing the side effects of radiation therapy. Once again, the two items are very similar, so she decides to save one of them for next year.

15. Which of the following are common side effects of external beam radiation therapy for the treatment of prostate cancer?
    1. penile edema
    2. proctitis
    3. cystourethritis

    A. 1 & 2 only
    B. 1 & 3 only
    C. 2 & 3 only *
    D. 1, 2, & 3

16. Which of the following is the LEAST common side effect when treating prostate cancer with external beam radiation therapy?
    A. penile edema *
    B. impotence
    C. diarrhea and abdominal cramping
    D. cystourethritis

Her watch chirps. It's 1:00 a.m. and she has completed 16 pretty good items. Ten of them can be used for this year's examination, and the remaining six can be used as replacement items when she covers this unit next year. She thinks back and remembers a few of the pointers that made writing them relatively painless.

First, Terri took some time to do some preliminary reading to identify the topics she wanted to cover. Nothing comes in handy like a good reference. Writing the objectives seemed to help, even though they weren't needed for every item.

Second, once started, she found that one item prompted another. So, she wrote multiple items on the same topics. And, although Terri didn't really follow the 5-step process recommended in the book, once she got started it really didn't matter.

Third, she remembered that along with the items that require basic recall, it's a good idea to mix in a few items that demand clinical reasoning and critical thinking skills. And finally, she used different formats, while writing only a limited number of negatively worded and combined-response items. As Terri turned out the light and closed her office door, she thought, "Hey, I bet I could do this for the ARRT."

Appendix D: Psychometrics Style Sheet

These are the editorial conventions used by ARRT's Psychometric Services Department in developing exams. Some of the conventions may seem arbitrary, but they are necessary for consistency across the certification categories.

GRAMMAR, USAGE, AND STYLE

PUNCTUATION

apostrophe

- There is no apostrophe used with plural forms of years: 1700s, 1940s
- There is no apostrophe used with plural numerals: count by 5s
- There is no apostrophe used with plural forms of acronyms: UVs, ICBMs, RVs

bulleted lists

- Lists must be appropriately punctuated. Lists comprised of one word or very short phrases do not require end punctuation. Lists comprised of complete sentences should end in periods. Lists comprised of a series of steps or related thoughts should end in semi-colons, with the final item ending in a period.

caps

- When negative words such as **NOT**, or **LEAST,** or other similar negatives are used in the stem, they should be in caps.

commas

- Use the serial comma at all times. **Example:** The flag is red, white, and blue.
- Use a comma in numerals of four or more digits (other than years). **Example:** 3,589
- Use commas (not semi-colons) followed by a single space to separate ordered pairs. **Example:** (2, 3), (5, 6), (8, 9)
- When numbers with multiple number places are spelled out within an item, do **not** insert commas. **Example:** two thousand five hundred seventy two.

contractions

- Do **not** use contractions in test items.

dashes

- An en dash (not a hyphen) is used to show continuity in numerals.
  Examples: January 11–16, pages 556–885, 1920–1945

dates

- Use an en dash to link two numerals that represent a continuous sequence.
  Examples: January 9–14, 1869–1875
- There is no apostrophe used with plural forms of years: 1700s, 1940s.

en dash

- An en dash (not a hyphen) is used to show continuity in numerals.
  Examples: January 11–16, pages 556–885, 1920–1945
- An en dash with spaces on either side is used for a subtraction sign.

hyphenation

- As a general rule, avoid using the hyphenation mode in text and items.
- Use hyphens for negative signs.
- For specific use of hyphenation relative to spelling, refer to the individual exam's conventions document.

- Compound adjectives – hyphenation rules are many and confusing. Generally speaking, when compound adjectives are shown hyphenated in a dictionary, one can assume that the expression is only hyphenated when it occurs directly before a noun:

Examples:

Hyphenated – "X-ray machine"
Not hyphenated – "the x rays were detected"

Hyphenated – "single-phase generator"
Not hyphenated – "operated on a single phase"

Use this	Not this
Nonionic	non-ionic
Revascularization	re-vascularization
Multidetector	multi-detector

Exception: intra-aortic NOT: intraaortic

- In general, do not use a hyphen to set off a prefix at the beginning of a word or a suffix at the end of a word.

initials

- Usage note: There are no spaces between the initials in proper names.
  Example: P.B.S. Pinchback

italics

- Use italics for the titles of books, newspapers, periodicals, government publications, etc.
- Use italics for lowercase letters denoting variables within Mathematics items. **Example:** $x - 3(2y + 7)$
- Do **not** use italics for uppercase letters used as labels within graphics.
- Use italics for *x* and *y* axes on graphs.

item punctuation

- In multiple-choice items with sentence fragments or single-phrase answer choices/options, the first word of each option is lowercase (unless it is a proper noun), and there is no terminal punctuation.
- In multiple-choice items that are sentence-completion items, the first word of each answer option is lowercase (unless it is a proper noun), and there is appropriate terminal punctuation.

question marks

- In multiple-choice items with quoted material at the end of the stem, be sure that any question mark related to the stem is placed **outside** the quotation marks.

semi-colons

- Lists comprised of a series of steps or related thoughts should end in semi-colons, with the final item ending in a period.

symbols

- Use hyphens for negative signs.
- Use en dash with spaces on either side for subtraction signs.
- Use appropriate accents in foreign words or names.
- Straight quotes denoting inches or pounds may be used if this is a concept being tested.
- Refer to specific conventions for special symbols such as beta (β).

CAPITALIZATION

proper names

- Proper names should be capitalized according to Dorland's Medical Dictionary.
- Eponyms: When an eponym is included in the name of a disease, syndrome, sign, position, or similar designation, capitalize the eponym but **not** the common noun. Consult Dorland's or Stedman's medical dictionaries. **Example:** Hodgkin's disease. Generally speaking, eponyms are written with an apostrophe; however, the medical profession is inconsistent when applying this guideline, so medical dictionaries such as Dorland's or Stedman's should be checked to verify specific spellings.

graphics/art labels

- When images contain capitalized labels or numbered parts, the answer choices should also contain the capitalized labels. **Examples:** Part 1, Number 2, Organ 4, Arrow 5, Box 7
- The decision to use numbers or letters may be driven by the original graphic. If the graphic came with labels, retain the use of whichever labels were on the original. If the image is "clean", either numbers or letters are acceptable. Numbers may be the preferred choice to avoid confusion with the answer choices; i.e., options A, B, C, and D.
- Refrain from using the pound sign (#) with numbers. Simply: What is arrow 5 pointing to?
- Initially, an attempt should be made to arrange any arrows in order; i.e., ascending either clockwise or counter clockwise; however, keep in mind that if additional arrows are needed in the future, it will be difficult to maintain the initial order.

NUMBERS

commas

- Use a comma in numerals of four or more digits (other than years). **Example:** 3,589
- Use commas (not semi-colons) followed by a single space to separate ordered pairs. **Example:** (2, 3), (5, 6), (8, 9)
- When numbers with multiple number places are spelled out within an item, do **not** insert commas. **Example:** two thousand five hundred seventy two

en dash

- An en dash (not a hyphen) is used to show continuity in numerals. Do not use the word "to" between numerals.
  Examples: January 11–16, pages 556–885, 1920–1945
- An en dash with spaces on either side is used for a subtraction sign.

italics

- Use italics for lowercase letters denoting variables within Mathematics items. **Example:** $x - 3(2y + 7)$
- Do **not** use italics for uppercase letters used as labels within graphics.
- Use italics for *x* and *y* axes on graphs.

data

- Refer to in plural form when appropriate. **Example:** these data

hyphenation

- Use hyphens for negative signs.

leading zero

- Use leading zeros with decimals to avoid confusion.
  Example: use 0.35 rather than .35
  use -0.35 rather than -.35

numbers/numerals

- Use numbers one through ten spelled out and numerals for 11 and up. When numbers are mixed, use numerals to express quantities.
- In mathematical questions, numerals may be preferable for numbers less than eleven to highlight key numbers. **Example:** A rectangle measures 4 in. by 6 in.

percent

- In text, spell out the word *percent*. **Examples:** one percent, 27 percent
- In charts, graphs, etc., use numerals and the percent symbol.
  Example: 10%

ranges

- When listing ranges, place the unit of measurement only at the end.
  Example: 25 – 30%

spacing

- Use a single space between numerals and symbols denoting operations (addition, subtraction, multiplication, or division).
- Use no space between numerals or letters/variables and associated properties: **Example:** 12^2 x^3 10_x

units of measurement

- Units of measurement associated with numbers should always be referenced in the stem and may be referenced in the options whenever possible.
  Example: What is the amount of separation in centimeters between the point A and point B?
  A. 2 cm
  B. 3 cm
  C. 4 cm
  D. 5 cm

Example:

m	meter	
ft	foot, feet	
in	inch	
L	liter	
sec, s	second	
hr	hours	
A	amperes	
mAs	milliampere seconds	
V	volt	
kVp	kilovolt peak	
Hz	hertz	
R	roentgen	(SI unit is C/kg)
rad	radiation absorbed dose	(SI unit is Gy)
rem	radiation equivalent man	(SI unit is Sv)
m	meter	
Ci	curie	(SI unit is Bq)
C/kg	Coulombs per kilogram	(conventional unit is R)
Gy	gray	(conventional unit is rad)
Sv	seivert	(conventional unit is rem)
Bq	becquerel	(conventional unit is Ci)

measurement

- It is acceptable to abbreviate units of measure (centimeters, inches, feet, kilograms, etc.) in stimuli, but these terms should be spelled out in the stems and answer choices unless otherwise indicated. Generally, if there are numbers (values) associated with the unit of measurement, the unit of measurement will be abbreviated. If the unit of measurement is used without a value, do not abbreviate.
- Use periods with abbreviations for units of measure that might be confused with an existing word (e.g., in.).
- Do **not** use periods for most measurement abbreviations (e.g., mm, cm, ml, kg, ft, mph, etc.) Follow AMA Manual of Style.
- For abbreviations specific to each modality, refer to that exam's conventions document.

staging

- Follow AMA Manual of Style for TNM staging abbreviation guidelines.

time

- Use a.m. rather than AM or A.M.

vertebrae and spinal nerves

- Follow AMA Manual of Style for abbreviation guidelines.

Examples:

Region	Vertebrae	Spinal Nerves
cervical	C1 through C7	C1 through C8
thoracic	T1 through T12	T1 through T12
lumbar	L1 through L5	L1 through L5
sacrum	S1 through S5	S1 through S5
coccyx	4 fused, not individually designated	coccygeal

Hyphens are used for intervertebral spaces (including neural foramina) and intervertebral disks as follows:

Space	Disk
C2-3 (space between C2 and C3)	C2-3 disk
T2-3 (space between T2 and T3)	T2-3 disk
L2-3 (space between L2 and L3)	L2-3 disk
C7-T1 (space between C7 and T1)	C7-T1 disk
L5-S1 (space between L5 and S1)	L5-S1 disk

Ranges of vertebrae are expressed as in the following examples; use letters for both the first and last vertebra in the indicated range:

Example: C3 through C7 3rd through 7th cervical vertebrae

ARRT

- The following list of abbreviations is specific to ARRT exams and may be added to.

General Abbreviations	HVL	half-value layer
	TLD	thermoluminescent dosimeter
	SI	system international
Positioning Terms and Abbreviations	AP	anteroposterior
	PA	posteroanterior
	LAO	left anterior oblique
	LPO	left posterior oblique
	RAO	right anterior oblique
	RPO	right posterior oblique
Position, Projection, and View	radiographic position	Refers to a specific body position and describes the patients physical position.
	radiographic projection	Refers to the path of the central ray.
	radiographic view	Refers to the patient's anatomy as seen by the film or other recording medium; used when describing a radiograph or other image.
Exposure Distances	OID	object-image receptor distance
	SID	source-image receptor distance
	SFD	source-film distance
	SSD	source-skin distance
	SAD	source-axis distance
	SCD	source-collimator distance
	SDD	source-diaphragm distance
	STD	source-tray distance
Units of Measure	m	meter
	ft	foot, feet
	in	inch
	L	liter
	sec, s	second
	hr	hours
	A	amperes
	mAs	milliampere seconds

	v	volt	
	kVp	kilovolt peak	
	Hz	hertz	
	R	roentgen	(SI unit is C/kg)
	rad	radiation absorbed dose	(SI unit is Gy)
	rem	radiation equivalent man	(SI unit is Sv)
	m	meter	
	Ci	curie	(SI unit is Bq)
	C/kg	Coulombs per kilogram	(conventional unit is R)
	Gy	gray	(conventional unit is rad)
	Sv	seivert	(conventional unit is rem)
	Bq	becquerel	(conventional unit is Ci)
Multipliers	μ	micro	μSv, μCi
	m	milli	mL, mrem, mA
	c	centi	cm, cGy
	k	kilo	kV, kVp
	M	million	MV
general rule		no periods; no spaces within a unit; no “s” for plural.	

PLURALS/POSSESSIVES

apostrophe

- Usage note: classmates’ opinions, people’s lives, and 1800s.
- There is no apostrophe used with plural forms of years: 1700s, 1940s
- There is no apostrophe used with plural numerals: count by 5s
- There is no apostrophe used with plural forms of acronyms: UVs, ICBMs, RVs
- Use the apostrophe with eponyms: Hodgkin’s disease

data

- Refer to in plural form when appropriate. **Example:** these data

SPELLING

conventions

- Consult either Dorland’s Medical Dictionary or Stedman’s Plus (our default spell checker).

drug names

- Consult the PDR for spelling of drug names.

GRAMMAR/USAGE

articles

- In multiple-choice items that use the sentence completion format, keep articles with the stem unless the answer options require different articles (e.g., "a" vs. "an").

negatives

- Avoid using a "double negative" in items.
  Example: All of [these] **except**...is **not**...
- In stems, capitalize and bold negative terms.

pronouns

- A pronoun must agree with its antecedent in number, gender, and person.
- In general, avoid use of pronouns to avoid confusion and reference to a specific gender.

parallel language

- Multiple-choice item answer choices/options **must** be syntactically parallel with the stem and with each other.
- Lists of items or topics must be parallel.

verbs

- A verb must agree with its subject in number and person.

TERMINOLOGY

drug names

- General rule: Use generic name, **not** brand name
- If the generic name is not common, then use the generic name followed by the brand name in parentheses. Check the PDR for correct spelling.

film

- Only use "film" if it is truly film (i.e., unexposed).

TEST ITEMS

"a (an)"

- Do **not** use this construction in test items.

"all of the above"

- Do **not** use this construction in test items.

"and/or"

- Do **not** use this construction in test items.

answer choices or options

- Answer choices are set in random order, not in order alphabetically or by length.
- When answer choices are numeric, they should be aligned by decimal point and placed in ascending order.
- For answer choices where numeric options are preceded by an additional designation (e.g., $), the designation should be aligned to the first digit of the options. **Example:**

A. $0.50
B. $0.75
C. $1.00
D. $1.25

combined-response items

- Items may contain more than one component or part. If so, these are labeled **1. 2. 3.** etc. The responses most commonly used by the ARRT are as follows:

Example:
A. 1 & 2 only
B. 1 & 3 only
C. 2 & 3 only
D. 1, 2, & 3

Or

Example:
A. 1 only
B. 2 only
C. 3 only
D. 1, 2, & 3

contractions

- Variations on these are acceptable, but should be discouraged.
- Do **not** use contractions in test items unless a contraction is being tested.

graphic

- Each graphic used in association with an item must have a label.
  Example: MAM23
- If a graphic is used in an item, refer to it in the item stem.
  Example: On MAM23, the arrow is pointing to....
- Refrain from naming the graphic. (e.g., On *image* MAM23, In *illustration* THR3, etc.)
- When referring to arrows, be consistent with the verb choice.
  Example: On MAM23, arrow 3 points to – or arrow 3 indicates.
- Do **not** use the pound sign # in either the stem or on the image to designate "number".

item formats

- **Closed stem:** This is usually a complete question. The answer choices do not have ending punctuation.
- **Open stem:** The stem is completed in the answer choices. The first word in the answer choices is lowercase and the answer choices do not have ending punctuation.
- Multiple-choice item answer choices/options must be syntactically parallel with the stem and with each other.

"is (are)"

- Do **not** use this construction in test items.

"none of the above"

- Do **not** use this construction in test items.

parallel language

- Multiple-choice item answer choices/options must be syntactically parallel with the stem and with each other.

3-option items

- The use of the 3-option multiple choice format should be reserved for items that clearly have only three choices — NOT in cases when a third distractor is difficult to ascertain.

Example: Following the administration of contrast, a patient's temperature will:

a. increase
b. decrease
c. remain constant

Appendix E
Bibliography

Books & Monographs on Item Writing and Test Development

Case, S.M. and Swanson, D.B. (1996). *Constructing Written Test Questions for the Basic and Clinical Sciences.* Philadelphia, PA: National Board of Medical Examiners (*or* www.nbme.org/about/publications.asp).

Haladyna, T.M. (1999) *Developing and Validating Multiple-Choice Test Items*. Hillsdale, NJ: Lawrence Erlbaum Assoc.

Hubbard, J.P. (1978). *Measuring Medical Education* (second ed.). Philadelphia, PA: Lea & Febiger.

Jacobs, L.C. & Chase, C.I. (1992). *Developing and Using Tests Effectively*. San Francisco, CA: Jossey-Bass.

Osterlind, S.J. (1998). *Constructing Test Items*. Norwell, MA: Kluwer Academic Publishers.

Raymond, M.R. (1992). *Guide to Test Item Development*. Kansas City, MO: American Nurses' Association.

Reid, J.B. (1990). *Item Writing Manual*. St. Paul, MN: American Registry of Radiologic Technologists.

Roid, G.H. & Haladyna, T.M. (1982). *A Technology for Test-Item Writing*. New York, NY: Academic Press.

Journal Articles About Item Writing and Test Development

Downing, S.M. (1992). True-false, alternate-choice, and multiple-choice items. *Educational Measurement: Issues and Practice*, *11*(3), 27-30.

Gronlund, N.E. (1985). Constructing objective tests of knowledge, in *Measurement and Evaluation in Teaching*, 5th ed., pp. 180-191. New York: Macmillan.

Haladyna, T.M. (1992). The effectiveness of several multiple-choice formats. *Applied Measurement in Education*, *5*, 73-78.

Haladyna, T.M. & Shindoll R.R. (1989). Item shells: a method for writing effective multiple-choice test items. *Evaluation & the Health Professions, 12*, 97-106.

Harasym, P.H., Price, P.G., Brant, R., Violato, C. & Lorscheider, F.L. (1992). Evaluation of negation in stems of multiple-choice items. *Evaluation & the Health Professions, 15*, 198-220.

Joorabchi, B. (1981). How to construct problem-solving MCQs. *Medical Teacher, 3*, 9-13.

LaDuca, A., Downing, S., & Henzel, T. (1995). Test development: Systematic item writing and test construction, in J.C. Impara (ed.), *Licensure Testing: Purposes, Procedures, and Practices*. Lincoln, NE: Buros Institute of Mental Measurement.

Millman, J. & Greene, J. (1989). The specification and development of tests of achievement and ability. In R.L. Linn (ed.) *Educational Measurement* (third ed), New York, NY: Macmillan.

Nitko, A.J. (1989). Designing tests that are integrated with instruction. In R.L. Linn (ed.) *Educational Measurement* (third ed.), New York, NY: Macmillan.

Page, G., Bordage, G., & Allen,T. (1995). Developing key feature problems and examinations to assess clinical decision-making skills. *Academic Medicine*, *70*(3), 194-201.

Shea, J. et. al. (1992). An adaptation of item modeling for developing test item banks. *Teaching and Learning in Medicine, 4*, 19-24.