Item Development Guidelines

with an Overview of the Examination Development Process

Table of Contents

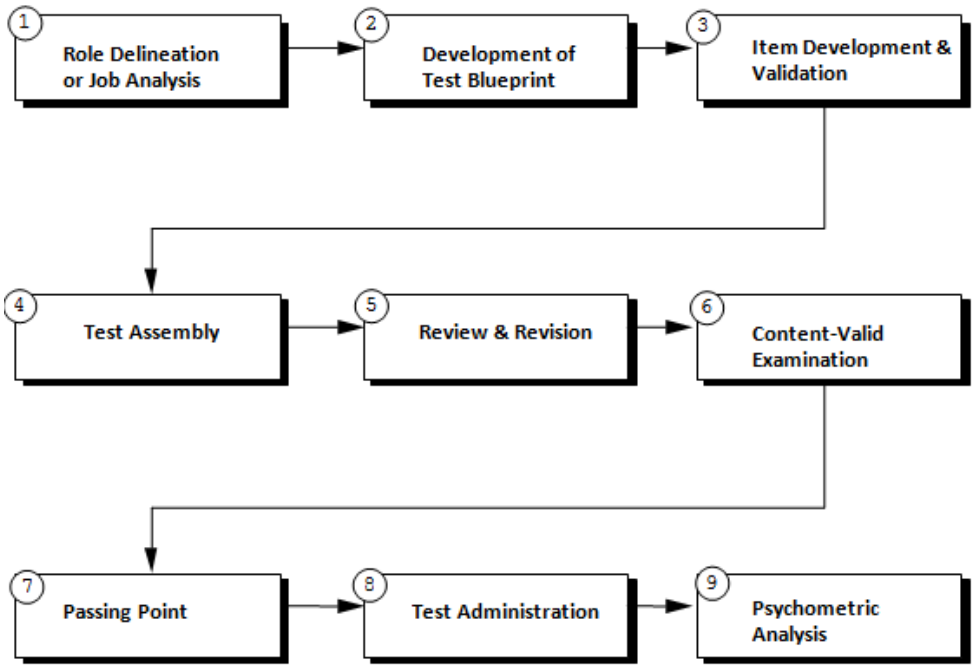
Introduction

Well-written items form the basis of a valid, reliable, and credible certification examination. High-quality items reflect the important skills and practices individuals use in their profession. This overview is presented to help the item writer understand some of the concerns in writing a high-quality item. The purpose of this guide is to provide:

(1) a description of each item type;
(2) general guidelines to help authors create valid test questions following the appropriate formats; and,
(3) basic examples of each question format, as well as pitfalls to avoid.

Overview of the Exam Development Process

As an author of test items, one should know how the writing task fits into the exam development process. Understanding the role and the significance of your participation allows for the best possible end product. The following diagram outlines the various steps in the exam development process and provides a visual workflow; the descriptions that follow discuss their importance and relevant activities encompassed within the steps.



Steps in the Development Process

Role Delineation or Job Task Analysis

Before development of an examination, a *professional practice analysis* (sometimes referred to as a *job task analysis* or *role delineation study*) is performed. Utilizing the expert knowledge of industry professionals, the study identifies domains of knowledge, tasks critical to the profession, and essential competencies requisite to a “minimally competent” professional.

The job responsibilities delineated by the subject-matter experts are arranged into performance domains, distinct tasks within those domains, and then further into skills and knowledge competencies required to perform those tasks successfully. These domains, tasks, and knowledge and/or skills are grouped and structured hierarchically. The domains, and sometimes tasks, are then validated through a survey of practitioners who review and rate the domains and tasks according to their importance, criticality, and frequency of performance.

Development of Exam Blueprint

In the next step, the results from the validation survey are used to develop a blueprint, or exam outline. This blueprint guides the item development and examination assembly process and ensures that the examination reflects the relative importance of the required knowledge and skills.

The numerical ratings associated with the importance, criticality, and frequency of each domain and task are translated directly into the percentage of items that should be included on the examination for a given content area (i.e., the more critical or frequently utilized a topic area is, the greater the number of questions associated with that topic that should appear on the exam).

An example test blueprint that includes domain and task weighting is shown below (*Note: this generic blueprint does not pertain to an actual examination*).

Domain	**010000**	**Anatomy, Physiology, and Pathophysiology**	**15%**
Task	010100	Respiratory	
Task	010200	Cardiovascular	
Task	010300	Nervous system	
Task	010400	Renal, gastrointestinal, and genitourinary	
Task	010500	Endocrine	
Domain	**020000**	**Pharmacology**	**25%**
Task	020100	General principles	
Task	020200	Inhalation agents	
Task	020300	Intravenous agents and reversal agents	
Task	020400	Local anesthetics	
Task	020500	Muscle relaxants and reversal agents	
Task	020600	Cardiac and vasoactive drugs	
Task	020700	Drugs acting on the CNS/ANS	
Task	020800	Adjunct drugs	
Domain	**030000**	**Basic Principles**	**35%**
Task	030100	Preoperative assessment/anesthetic plans	
Task	030200	Monitoring	
Task	030300	Positioning	
Task	030400	Fluid therapy	
Task	030500	Perioperative complications	
Task	030600	Airway management	
Domain	**040000**	**Advanced Principles**	**25%**
Task	040100	Regional anesthesia	
Task	040200	Pediatric anesthesia	
Task	040300	Obstetrical anesthesia	
Task	040400	Geriatric anesthesia	
Task	040500	Management of surgical specialty procedures	

In the example blueprint, there are 4 major topic areas, or domains. Each domain is further broken into subcategories called tasks. Each domain may have several associated tasks. The tasks may even be further subdivided to recognize specific knowledge or skills required within the task.

The weighting shown on the example blueprint indicates that no matter how many items make up the final exam, 15% of the items must be classified to Domain 1 or one of its subcategories. Likewise, 25% of the items must be classified to Domain 2 or one of its subcategories.

Classification of Items

All examination items are developed in association with the test blueprint. When classifying items, the first 2 digits of the classification code represent the highest level of the blueprint hierarchy: a domain being tested. The second 2 digits represent the first subsequent level of a domain: typically the task statements being tested. The third pair of digits represents the next subsequent hierarchy level being tested. This continues to the lowest level of classification that the blueprint defines.

Providing correct classifications is essential when writing items. Item classifications enable test developers to easily identify various types of questions and ensure that the entire domain of knowledge is represented on the examination.

Example: Classification number 010200 would represent Domain 1, Task 2.

Every item must reflect a single specific piece of knowledge or skill, as required in the test specifications. The courts of the United States require that all items be explicitly linked to a facet of professional performance. We make this link between item and performance using the classification system. An examination containing even 1 item that is not classified is not legally defensible.

Validation

Each question is reviewed and validated by subject-matter experts and must have a verifiable reference. Each item is classified by content category and validated according to its appropriateness to the relevant practitioner population (in this case, nurse anesthetists). After development, items are reviewed to ensure that they are psychometrically sound and grammatically correct.

Exam Assembly

Each examination is drafted by selecting the appropriate number of items from each content area, as specified in the test blueprint. The items included in the selection are reviewed by the test developer using data, when available, collected during development steps and/or from administrations to complement the following questions:

- Do the items fulfill the exam blueprint?
- Do the items have content that is relevant, frequently used, and critical to job performance?
- Do the items refrain from providing answers to other test questions?
- If available, are item statistics appropriate for the examination?

Exam Review and Revision

Although each question is written by an expert in the profession, it is important that each item be reviewed and validated by other experts in the profession. This process helps to ensure that each item is appropriate for the certification examination and free from individual biases.

The draft examination is reviewed by subject-matter experts for technical accuracy and by psychometric experts to ensure its psychometric integrity. Item performance data, which may be available if items have been used on previous examination versions, is reviewed by the subject-matter experts. Using the statistical item analyses, inappropriate or questionable items are either revised or omitted from the examination.

When reviewing and validating examination questions, experts are asked to consider questions such as:

- Is the knowledge tested by the item important to the assessment of competent practice?
- Does a correct response to the item differentiate adequate from inadequate performance for the practitioner?
- Does the item have a verified reference?
- Is the item appropriate for the relevant practitioner population?
- Is the keyed answer correct?
- Can the keyed answer be defended if necessary?
- Are the distractors incorrect but still plausible?

Content-Valid Examination

The procedures described above are accepted procedures for developing reliable and content-valid examinations. Each step in the test construction process is carefully documented. Multiple reviews by content and psychometric experts and the use of stringent criteria strengthen the validity of the test. Continuous evaluation of each examination's reliability maintains the consistency of the test to measure examinees' skills accurately.

Standard Setting and Passing Point

The cutoff score that separates examinees who pass a credentialing examination from those who fail must be based on the minimum competency level required to protect the public from harm. Documentation of the standard-setting procedures is a key component for establishing validity of a credentialing examination. A valid credentialing examination must have a defensible passing score.

NBCRNA Standard-Setting Process

Periodic standard-setting studies for examinations are generally regarded as a best practice in credentialing. The frequency with which these studies are conducted should correspond to how quickly a given profession is perceived to change. Generally speaking, it is recommended that testing programs undertake a standard-setting study, following a comprehensive professional practice analysis, every 4-5 years.

In most cases a panel consisting primarily of subject-matter experts is assembled to provide guidance on the level of knowledge that must be demonstrated to be considered competent to practice safely. The panel members are selected to reflect the diversity of the profession, including gender and ethnicity, practice setting, geographic region, professional experience, educator versus practitioner status, etc.

Using one of a number of methods for setting performance standards, the panel arrives at a single consensus point on the ability continuum and recommends a pass/fail point on the test. The panel is typically convened to review the standard-setting methodology, become familiar with the purpose of the test, discuss in depth the knowledge, skills, and abilities required of an entry-level practitioner, and

make the judgments that will be used in determining the passing score for the exam. The panel's passing score is subject to review, adjustment, and ratification by the governing certification body.

Test Administration

Test administration procedures must ensure consistent, comfortable testing conditions for all examinees. For secure examinations, procedures must address examinee admission into the room, seating charts, display of information signs, security, time allocation, and other aspects of the administration. Testing facilities must meet guidelines that ensure security, proper room size, ventilation, restroom facilities, accessibility for those with disabilities, and noise control.

Psychometric Analysis

Following the test administration, item statistics are reviewed to ensure quality and validity. Statistics that are evaluated include the item difficulty and the item discrimination. Items with poor performance statistics are evaluated by subject-matter experts prior to scoring. These items are then tagged for review at the next meeting. Further information regarding statistical analysis can be found in the appendix of this guide.

Guidelines for Developing Multiple-Choice Items

Types of Multiple-Choice Questions

All NBCRNA examinations use multiple-choice items. There are 2 basic types of multiple-choice questions: *interrogative* and *completion*. They can be used to assess candidates' recall of factual knowledge, or their competencies for application and analysis of information. Item writers should consider whether an interrogative or completion question form will be clearest to the candidate.

Interrogative Questions

In interrogative multiple-choice questions, the *stem* (the body of the question including any scenarios or qualifying information) states a complete problem in the form of a question. The options, although plausible, contain only 1 correct response and 3 definitely incorrect distractors. Questions frequently include the phrase *which of the following* to direct the candidate to the consideration of the listed options rather than the universe of possible answers. The *which of the following* construction may be omitted and replaced with *what* when the question calls for simple factual recall. An example of a recall interrogative question is:

What city is the capital of Florida?
A. Jacksonville
B. Tallahassee
C. Miami
D. Tampa

Sometimes the correct response must be the best of the alternatives, which are distractors that are demonstrably less appropriate. An example of the best-answer format for multiple-choice questions is:

*Which is the **MOST** effective tool to drive a nail into a wall?*
A. Hammer
B. Pipe wrench
C. Brick
D. Textbook

With questions such as the above, writers should make sure the stem confines the choices to the available options. For example, test takers are instructed to select the *best* or *most appropriate* or *most complete* of the response alternatives.

Completion

Other multiple-choice questions are better written as incomplete sentences. The stem poses a problem, and allows the candidate to complete the sentence by selecting the correct option. Completion-type multiple-choice items should **NOT** use blanks (e.g., *A _____ is a type of evergreen tree*.). **The completion should always require the test taker to finish the sentence**. An example of a completion question is:

The capital of Florida is:
A. Jacksonville.
B. Tallahassee.
C. Miami.
D. Tampa.

Research findings favor **interrogative over completion** form—use completion only when clarity requires it.

Stem Construction

Guidelines

➡**Maintain focus**. A well-constructed stem for a test item is one that could be answered without seeing the options. The intent of the stem is to focus the test taker directly on the tested information. The following is an unfocused question:

Which of the following is true about musical instruments?
A. Guitars usually have 6 strings.
B. Grand pianos have 67 keys.
C. Clarinets are the largest woodwinds.
D. A sousaphone is another name for a tuba.

In the example, the examinee cannot understand the objective of the question without reading all the options. A focused way of testing the same knowledge might be:

*How many strings does a guitar **MOST** often have?*
A. 4
B. 5
C. 6
D. 12

Written this way, the information that the examinee is asked to provide is clear, and confounding questions of reading comprehension are not involved.

➡**Avoid negative construction in stems**. That is, avoid using the words NO, NOT, LEAST, EXCEPT, etc.

*All of the following are high glycemic index foods **EXCEPT**:*
A. Rice
B. Grape juice
C. Onions
D. Potatoes

Again, the examinee's logical reasoning, reading comprehension, and language skills become confounders in the inferences we can make about his/her competence. The same topic might be handled better in a multi-select question (see section on Multi-Select items below), or as follows:

*Which food has the **LOWEST** glycemic index?*
A. Onions
B. Rice
C. Potatoes
D. Grape juice

➡**Write questions that are clinically relevant**. However, in including patient cases, while some authentic, potentially confusing details can be appropriate because many professional situations can be unclear on first encounter, bear in mind that the candidate is under time pressure and cannot consult colleagues and references.

➡**Include the central idea and most of the phrasing in the stem**. In the following example, some repeated information impedes readability:

*What are the 2 nonmodifiable risk factors **MOST** associated with arthritis?*
A. Age and family history
B. Age and history of infection
C. Age and gender
D. Age and ethnicity

The question could be rewritten with the repeated information in the stem:

*Besides age, what nonmodifiable risk factor is **MOST** associated with arthritis?*
A. Family history
B. History of infection
C. Gender
D. Ethnicity

➡**Teaching** is the inclusion of fact statements in an item leading up to a question, and should be avoided. Examinees should come to the test with all the information they need to answer any question. If extra information is needed to answer a question, it may not be a good item for testing. An example of a question with teaching in the first sentence is:

Mivacurium is a nondepolarizing neuromuscular blocker. The primary mechanism of action of mivacurium is via what neurotransmitters?
A. Acetylcholine
B. GABA
C. NMDA
D. Dopamine

In this example, the question would test the same information with the first sentence deleted.

➡**Items should be constructed as clearly and directly as possible.** Use as few words as possible to adequately convey the question's meaning. Generally, the fewer words used, the clearer the questions. Example:

Original stem:
Which position change would you choose to attempt to improve hemodynamics?

Revised stem:
Which position change improves hemodynamics?

➡**Ask about only 1 concept in a question.** Compound questions such as the following should be separated into 2 questions:

The half-life and duration of action of pancuronium in minutes are, respectively:
A. 45 and 90
B. 60 and 120
C. 75 and 150
D. 90 and 180

The same applies to the following:

Original stem:
What shape or angle should a stylet be for nasal intubation?

Revised stem:
Nasal intubation requires a stylet of what shape?

➡**Include only relevant information**. Time limitations and fairness do not permit distracting details.

➡**Refrain from using the phrase *which of the following* in the stem**. Instead, simply tell the reader what "the following" are. (e.g., *Which types of airway are appropriate for a patient in status asthmaticus?*)

➡**Words in the stem should correspond to the options.** For example, the word "statements" is used in the following stem to describe the options, but the options are not statements per se; they are descriptions of medical situations. Example:

Original stem and options:

Which of the following statements for explaining the rationale for decreased epidural local anesthetic requirements in a healthy parturient is correct?
A. Decreased ability to metabolize local anesthetics
B. Decreased volume of the epidural space
C. Increased neuronal sensitivity to local anesthetics
D. Venous engorgement of the epidural space

Revised, streamlined stem:

Which situation is a rationale for decreasing an epidural local anesthetic in a healthy parturient?
A. Decreased ability to metabolize local anesthetics
B. Decreased volume of the epidural space
C. Increased neuronal sensitivity to local anesthetics
D. Venous engorgement of the epidural space

➡**State lapses of time in exact measurements**. Use "4 hours," rather than writing "a few hours." Also, always include the **appropriate units** for any measurement:

Original: *A patient with head trauma has a heart rate of 160, blood pressure of 90/40, and respiratory rate of 40.*

Revised: *A patient with head trauma has a heart rate of 160/min, blood pressure of 90/40 mm Hg, and respiratory rate of 40/min.*

➡**The stem should not be written in the second person** (i.e., what would *you* do if . . .). Technically, the use of *you* converts the item from a test question into a survey question, and any response the examinee chooses is correct.

➡**Avoid noncommittal wording**. This includes may/may be, might/might be, can/can be, could/could be, and so forth. These words denote that the supposedly correct answer is in some cases untrue, which could confuse the examinee. If the examinee is able to justify a single exception to the statement in the stem, the question may have multiple correct answers. Examples:

Original stem:
At what postconceptual threshold age could an infant be considered at minimal risk for developing postanesthesia apnea?

Revised stem:
At what threshold age is an infant at minimal risk for postanesthesia apnea?

And . . .

Original stem:
Pickwickian syndrome may be characterized by:

Revised stem:
Pickwickian syndrome is characterized by:

And . . .

Original stem:
Obstructive sleep apnea may be associated with:

Revised stem:
Obstructive sleep apnea is associated with:

➡**Be careful using superlative qualifiers (e.g., MOST, LEAST, BEST, etc.) in the stem.** A stem constructed around one of these words may be subjective. Additionally, the key must be absolutely clearly superior to the distractors. If there is no alternative construction, the qualifying word should be written uppercase and boldface. Consider the item:

*Which opioid is **MOST** likely to produce adverse effects when administered intrathecally?*
A. Fentanyl
B. Sufentanil
C. Morphine
D. Meperidine

Since the question is not asking for the only drug that produces side effects, an uppercased and boldfaced **MOST** helps clarify the question while giving nothing away. Even though a careful reading of the stem seems to eliminate the need to emphasize *most*, the question's intent is made clear for examinees (and writers!) by uppercasing and boldfacing.

A similar example is the stem:

The major risk factor for nerve injury during xyz surgery is:

This stem could be read as saying that *ALL* of the options are risk factors for nerve injury, but that the correct answer represents the riskiest of them. But the writer's intention is unclear. If the writer wanted the candidate to pick the riskiest factor out of 4, (s)he could instead write:

*Which situation puts a patient **MOST** at risk for nerve damage during xyz surgery?*

However, if the writer wants the candidate to pick the **1** situation that presents a risk factor, and the distractors are not risk factors, the author can eliminate the need to use the qualifier and write:

A risk factor for nerve injury during xyz surgery is:

➡**Eliminate cues**. If the stem asks the candidate to identify, for example, a risk factor, the key should not contain the words *dangerous*, *harmful*, *questionable*, *hazardous*, etc. Conversely, the distractors should not contain the words *neutral*, *benign*, *lack of*, *avoidance of*, *prevention of*, or other words that typically indicate harmlessness or promotion of safety. Also, if a word in the stem is repeated in an option, it could cue the test taker to—correctly or wrongly—pick it. For example:

The intraoperative monitor that provides the most immediate evidence of an evolving laryngospasm in a child is a(n):
A. pulse oximeter.
B. end-tidal carbon dioxide monitor.
C. precordial stethoscope.
D. peak inspiratory pressure reading.

Option B could be changed to end-tidal carbon dioxide *indicator* to prevent it from linking to the stem.

Another example of cuing is when one item cues another. A very basic but extremely problematic example of this is including the following 2 items on the same test:

Chicago is adjacent to which Great Lake?
A. Ontario
B. Huron
C. Michigan
D. Erie

And . . .

Which of the following is a Great Lake?
A. Winnebago
B. Louise
C. Erie
D. Placid

Notice how the first question provides the answer to the second. These items are referred to as *enemy* items to capture this cuing relationship.

➡**Do not overuse shorthand.** If in doubt, spell it out. High-stakes exams should not look like alphabet soup. Please write in complete sentences, complete with "little" words such as articles (*a, an, the*), prepositions, and conjunctions. Avoid using abbreviations except in approved instances. Examples:

Original: *BP 140/102 mm Hg*
Revised: *a blood pressure of 140/102 mm Hg*

And . . .

Original: an *HR of 60/min*
Revised: *a heart rate of 60/min*

➡**Allow time for editing and other types of item revisions.** The best items are improved through multiple reviews.

➡**If possible, include more than one published reference for each question.** An examination containing 1 unreferenced item is legally indefensible.

Response Option Construction

Guidelines

➡**The key consideration in writing response options is that *distractors*, or incorrect response options, must be plausible.**

- Think of common misconceptions and common errors made by professionals in the field. Use familiar yet incorrect distractors.
- Be careful not to cue test-wise examinees with word associations between the stem and correct response, absurd distractors, formal prompts, or overly specific or overly general options.
- Phrase options positively (i.e., avoid *not*).
- Keep all response options **grammatically consistent** with the stem. Test-wise candidates assume that the correct response will be grammatically consistent, so they eliminate inconsistent options quickly.
- Keep all options in an item **consistent in content** (i.e., all diagnoses, all nerves, etc.).
- Keep the **length of the options** fairly consistent.
- Keep options independent; options should not **overlap** conceptually or numerically.
- Avoid **absolute terms**: *always, never, all*, or *only* tend to be perceived by test takers as associated with false content.
- Avoid using regional differences in practice as options.

➡**Place options in logical or numerical order**. For ease of reading and to make options appear equally attractive, if the responses are 1, 3, 6, and 9, list them that way, or the reverse. Order non-numerical responses alphabetically or logically, as below:

The musculocutaneous nerve arises from which nerve roots?
A. C4
B. C5
C. T1
D. T2

➡**Options should be grammatically parallel**. If options stand out grammatically, test-wise candidates can eliminate them quickly. If 3 options are nouns, the fourth should also be a noun; if 3 options start with a verb, so should the fourth, and so on.

Example 1: Option C does not conform to the construction since it is a phrase rather than a single word.

A. Bupivacaine
B. Procaine
C. Administration of tetracaine
D. Lidocaine

Example 2: Option C stands out because of its differing construction.

A. Increased sympathetic tone
B. Increased conversion of angiotensin I to angiotensin II
C. The high osmolarity of the injectate
D. Increased constriction of capacitance vessels

Example 3: Option A differs in that it isn't given as a percentage.

A. Normosol
B. 5% dextrose
C. 0.9% sodium chloride
D. 10% dextran 40

Example 4: Option C differs in that it is a sentence whereas the other options are phrases.

A. Presence of unidirectional valves	[phrase]
B. Lack of soda-lime carbon dioxide absorption	[phrase]
C. Rebreathing is independent of fresh gas flow	[complete sentence]
D. High amount of resistance	[phrase]

Example 5: Here, a gerund (-ing noun) is used in 3 options (A, B, D), so it should be used in the fourth as well.

A. Elevating the head of the bed
B. Using slight reverse Trendelenburg
C. Flexion of the hip and knee joints (Should be: Flexing the hip and knee joints)
D. Dropping the foot 30°

➡**Avoid using paired option constructions**. Example:

A. An increase in end-tidal flow
B. An increase in respiratory rate
C. A decrease in end-tidal flow
D. A decrease in respiratory rate

The first and third options have the same construction as one another, and the second and fourth options have the same construction as one another; in these the candidate would know that one of each pair was incorrect, thus increasing his or her chances of guessing correctly.

➡**Where appropriate, use *a, an*, or *a(n)*** at the end of a stem to avoid repeating these words at the beginning of each option. Example:

In xyz disease, the patient should be examined for the presence of a(n):
A. ulcer.
B. tumor.
C. hernia.
D. cyst.

➡**Do not use *all of the above* and *none of the above* as distractors.** In *all of the above* questions, if a candidate can eliminate 1 of the other 3 distractors, then *all of the above* is also eliminated. Thus, the chances of the candidate guessing the correct response are substantially improved.

With *none of the above*, an examinee can answer correctly by knowing 3 of the options are incorrect, but does not thereby make a positive demonstration of knowledge of what is being tested. For example:

The capital of Illinois is:
A. Rockford
B. Champaign
C. Peoria
D. None of the above

A candidate answering D is correct, but at no time proves that he/she knows Springfield is the capital of Illinois. In fact, the candidate may think the capital is Chicago, so inferring examinee competence from the answer would be a mistake. Imagine the same sort of error in a medication context:

The most appropriate drug to give a patient with history x *is:*
A. Drug A
B. Drug B
C. Drug C
D. None of the above

The examinee may answer correctly, but still believe the correct drug is one that would actually be lethal in the context.

➡**Every question written must have the correct number of response options.** Before writing questions, make sure you are aware of the number of response options required by the certification program for which you are writing.

Rules for Using Graphics and Attachments

Item writers are strongly encouraged to use graphics as stimuli for any of the item types. When using graphics, please adhere to the following rules:

➡**Item writers may use graphics from any of the following sources**:
- Actual patients and cases, provided necessary consents have been supplied with question submission
- Equipment and instrumentation displays
- Simulation lab prints and pictures
- Cadaver lab photos
- Images from NBCRNA's contracted medical illustrator
- Public domain (Wikimedia, *Gray's Anatomy*, some government sources), with citation
- Nagelhout and Plaus, fifth edition—other textbook graphics will require express permission from the publisher or copyright holder.
- Other resources that have been prior-approved by the NBCRNA
- Websites or other sources provided proper written permission has been received.

➡**Avoid "window-dressing."** Do not put a graphic in a question merely for the sake of adding a graphic. If you are including graphics as part of the stem, make sure the graphic is pertinent and necessary to answering the question. In other words, it should be essential to refer to the graphic to answer the question.

➡**Do not ask items that require test takers to simply look up information in the material provided**; ask test takers to make inferences and decisions.

➡**Multiple items on a test form can use the same attachments.** It saves reading time, but items **MUST** be independent (i.e., you do not need to answer one item correctly in order to answer another item correctly).

➡**Do NOT use drawings or photographs that are copyrighted by another individual or organization, unless written permission from the source has been procured.**

➡**The preferred format for graphics files is jpg**. Please submit all graphics in this format, with file size below 2 MB.

➡**Some graphics may need modification** to remove (or insert) labels, arrows, etc. To the extent possible, item authors should perform any **photo editing** themselves. Most modifications can be performed using Microsoft Paint, a free program in Windows. If item writers cannot perform the modifications, then detailed instructions for the changes should be documented in the Notes or Comments fields of the item.

➡**We currently have the capability to make graphical response options** as well as graphical stems.

Guidelines for Developing Alternate Item Types

Multi-Select

Description

This question format is similar to a multiple-choice question, consisting of a stem (incomplete statement or question) followed by options from which the examinee must select the correct responses. However, in contrast to the multiple-choice format, which consists of a stem and 4 response options (only 1 of which is correct), multi-select items consist of a stem and 4 to 8 response options, and more than just 1 option is a correct response. The question stem will indicate how many responses are correct. The examinee must select *all* the correct responses in order to be awarded credit. Example:

What are the hemodynamic goals of hypertrophic cardiomyopathy?
(Select 2.)

A. Decrease contractility
B. Decrease preload
C. Increase afterload
D. Increase heart rate

Guidelines for Stem Construction

In general, follow the same guidelines you would follow for **constructing a stem** for a multiple-choice question. (See the *Guidelines for Developing Multiple-Choice Items* section of this document):

➡**Generally, the total number of response options should be twice the number of correct responses.** For example, if there are 3 correct responses to the stem, then try to write 6 total response options. Multi-select questions should include 4 to 8 response options.

➡**After the stem, indicate the number of correct responses.** Specifically, provide the instruction, "**Select 2/3/4/etc**." Examples:

What are potential complications of pulmonary artery catheter insertion?
(Select 2.)

A. Cardiac perforation
B. Left bundle-branch block
C. Mitral valve rupture
D. Pulmonary infarction

Characteristics of Eaton-Lambert Syndrome include:
(Select 3.)

A. positive response to anticholinesterase agents.
B. improved strength with activity.
C. reduced acetylcholine release.
D. destruction of acetylcholine receptors.
E. postjunctional defect.
F. sensitivity to all muscle relaxants.

➡**Stem content should focus on a single theme or problem. Focus** is essential to writing clear test items. A lack of focus can contribute to confusion on the part of the examinee.

➡**The stem should be economically worded.** Avoid unnecessary verbiage.

➡**The stem should be grammatically correct**, both alone and in conjunction with the responses.

➡**The stem should not be written in the second person** (i.e., what would *you* do if . . .). Technically, the use of *you* converts the item from a test question into a survey question, and any response the examinee chooses is correct.

➡**The multi-select format should essentially eliminate any need for negatively phrased stems** (e.g., "which of the following is NOT . . . " or "all of the following EXCEPT . . . "). The following is a negatively phrased question:

Which drug is NOT associated with methemoglobinemia?
A. Benzocaine
B. Ketamine
C. Phenytoin
D. Metoclopramide

This could be rewritten using multi-select as:

Which drugs are associated with methemoglobinemia? ***(Select 3.)***
A. Benzocaine
B. Ketamine
C. Phenytoin
D. Metoclopramide

(Note that in the rewrite, some new, false distractors should be added.)

Guidelines for Response Construction

Also follow the same guidelines you would follow for **constructing responses** for a multiple-choice question. (See the *Guidelines for Developing Multiple-Choice Items* section of this document.):

➡**Response options should be homogeneous**, or conceptually related to each other, and limited to 1 topic.

➡**Response options should fit grammatically with the stem** if the item form is sentence completion.

➡**Responses should be parallel in grammar, sentence structure, and length.**

➡**Distractors (incorrect responses) should be objectively incorrect, but plausible** without being tricky. Possible sources for distractors may be common misconceptions.

➡**If the responses are numbers, they should be arranged in numerical order.**

Drag-and-Drop

Description

These questions involve clicking and dragging objects to corresponding targets and may take the form of matching or placing objects in order.

Guidelines

➡**Always use the following rule: *Shorter on the left, longer on the right*.** The left/right (shorter/longer) structuring should be considered when creating both the stem and the options for drag-and-drop items.

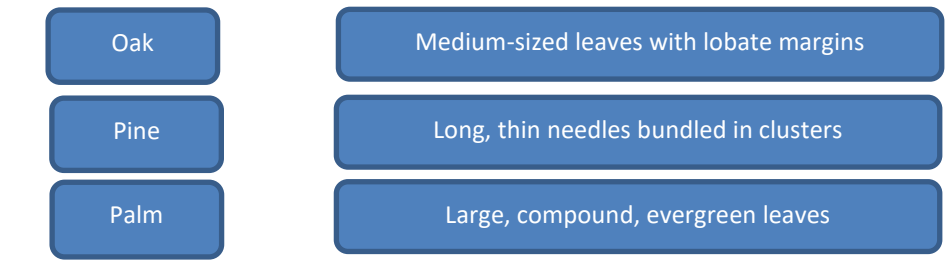
➡**Stems should use the following language:**

Match each [left—response] *with each* [right—target]

For instance, a test taker should not be asked to match a tree to a description of its leaves. A tree is a singular object and its name is likely to be shorter than the description of the named component.

NO: *Match each description of leaves with the tree that bears them.*
YES: *Match each tree with the description of leaves that it bears.*

➡**Options should accordingly be structured in the following manner:**



Drag-and-Drop Matching

Adhere to the following guidelines when organizing *responses* (left column) and *targets* (right column):

➡**Use singular nouns** in the stem of an item asking the candidate to match single options. Example:

Original stem: *Match the interventions with their mechanisms*.
Revised stem: *Match each intervention with its mechanism of action.*

➡**Identify a theme, a general concept, or a genre that ties all of the matchable options (responses and targets) together (e.g., diseases, drug classes).**

➡**Limit matching to 3 or 4 matchable options.**

➡**Keep the response text (left column) to single words or short phrases.** Limit the target text (right column) options to 7 or 8 words.

➡**A response should be matched to each target, with no unmatched options.**

➡**Some matching questions take the form of causes and effects. For example:**

Match the disease process to the set of symptoms.

Place causes (disease processes, in this case) in the left column and effects (symptoms) in the right column.

Drag-and-Drop Ordering

Adhere to the following guidelines when organizing *responses* (left column) and *targets* (right column):

➡**In writing the stem or lead-in statement, specify how the objects should be ordered.** Example:

Original stem: Order the following objects by their specific mass.
Revised stem: Order the following objects by their specific mass, from least to greatest.

➡**Place responses to be ordered in the left column.** The right-column targets should read, “First,” “Second,” “Third,” “Fourth,” etc.

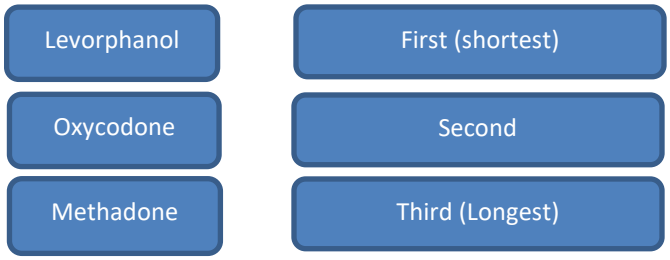
➡**Limit the objects to be sorted to 3 to 4 objects.**

➡**A response should be matched to each target with no unmatched options.**

➡**Keep the response text (left column) to single words or short phrases.** Limit the target text (right column) options to 7 or 8 words.

Example: Drag-and-Drop Ordering[1]

Place the opioid drugs in order of the duration of their half-lives, from shortest to longest.



[1] Nagelhout & Plaus. *Nurse Anesthesia.* Fourth edition. Philadelphia: Saunders, Elsevier; 2009: 1258.

Short Answer

Description

In this question format, examinees are asked to respond by typing in a numerical response, typically a whole number (no decimals) or a number with 1 or 2 digits right of the decimal point. The question will indicate to the examinee the required format of the response.

Guidelines

➡**Construct questions with <u>numerical responses only</u>.**

➡**Be very specific about how the answer should be supplied (whole number, number of decimal places, measurement units).** Closely follow the model text provided in the examples below:

> *Enter your answer below as a whole number with no decimal places, in liters per minute.*
> *Enter your answer below as a whole number with no decimal places and no units.*
> *Enter your answer below as a number with 2 decimal places, in mm Hg.*
> *Enter your answer below as a number with 1 decimal place, in liters.*
> *Enter your answer below as a whole number with no decimal places, in percent.*
> *Enter your answer below as a number with 1 decimal place, in volumes percent.*

➡**If a calculator is likely to be needed, indicate somewhere after the text of the stem that one is available onscreen by clicking the calculator link.**

➡**If the correct answer to the calculation question involves rounding, be sure to indicate the key such that the rounding is performed as the final step**. In other words, do not round intermediate values. (The instruction to round the answer as the final step of the calculation is provided in the examinee materials.)

➡**In general, calculation questions (involving well-known formulas) work better than questions involving ranges of approved values.**

➡**If the answer to the question is a range of acceptable values, please provide all acceptable answers in the item key**. For example, if the correct answer to a question is 20-30 mL, you must specify all whole number values between 20 and 30 (e.g., 20, 21, 22, 23, . . . , 29, 30). Also, if the answer is a range of possible values, please provide 2 textbook references documenting the correct range.

➡I**f the answer to the question involves a computation using a formula, provide in the Notes or Comments field of the item template an explanation of how the correct answer is obtained.**

➡**Consider whether a particular calculation may have more than 1 acceptable formula** for arriving at an acceptable answer. In this case, all answers reached by any of the acceptable formulae should be included in the answer key.

➡**Likewise, be aware that considerable variation may exist in drug dosages** that could be considered clinically acceptable, even if one text cites a single dosage. The solution to dosage items should be found consistently and discretely among a number of textbooks. If dosages vary among texts, the answer key should include all acceptable values, and the question should include a text reference for each acceptable value.

➡**If it is possible, consider "cloning" items** which involve calculations by simply changing the factors used to perform the calculation. For example:

Original stem:
A patient is 72 inches tall and weighs 200 pounds. Calculate this patient's body mass index (BMI).
Enter your answer below as a whole number with no decimal places, in kg/m^2.

Cloned stem:
A patient is 80 inches tall and weighs 250 pounds. Calculate this patient's body mass index (BMI).
Enter your answer below as a whole number with no decimal places, in kg/m^2.

Additional Examples

Calculate cardiac output given the following hemodynamic parameters: stroke volume, 60 mL; blood pressure, 150/70 mm Hg; heart rate, 50/min.
Enter your answer below as a whole number with no decimal places, in L/min.

_____ L/min

Use of positive pressure greater than how many mm Hg with a laryngeal mask airway may cause stomach inflation?
Enter your answer below as a whole number with no decimal places, in mm Hg.

_____ mm Hg

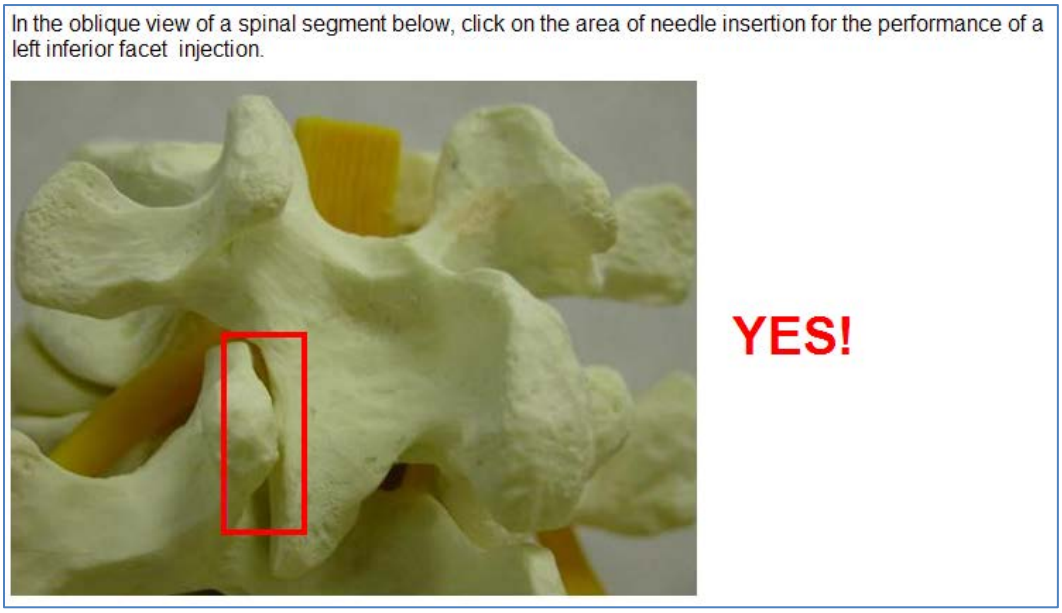
Hotspot

Description

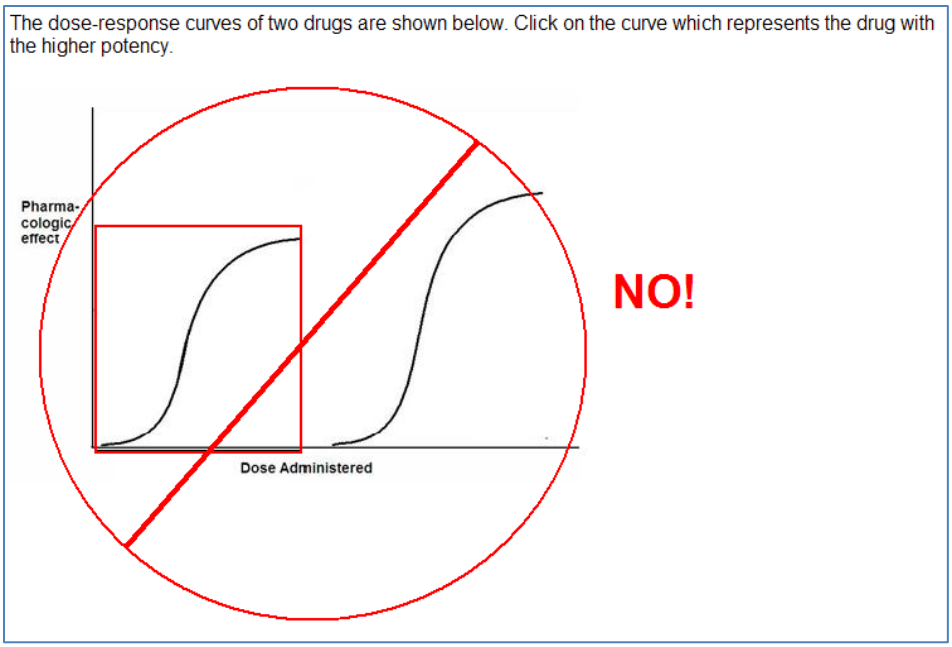
In this question format, the examinee is presented with an image or a figure. The examinee indicates the answer by clicking on a region of the presented image or figure. An X appears at the location of the click. To receive credit, the examinee must place the X inside the region of the image determined as correct by subject-matter experts.

Guidelines

➡**The hotspot format is best suited to questions that elicit many (or even infinite) potential places to click.** By contrast, if the question is formulated to allow only 3 or 4 possible click-points, then it might as well be a multiple-choice question. A question with many possible places to click is shown below.



In contrast, the example below really only allows for 2 plausible points to click and could be better reformulated as a multiple-choice question.



➡**All hotspot items are graphics-based.** Please see the guidelines in this manual for using images.

➡**Use the words *diagram, illustration, photo, graph*, *image,* etc.** Avoid referring to an image as a picture or figure. State in the stem what the image is. Examples:

In the photo of the inner arm, click on the area . . .
In the illustration, which waveform represents . . .

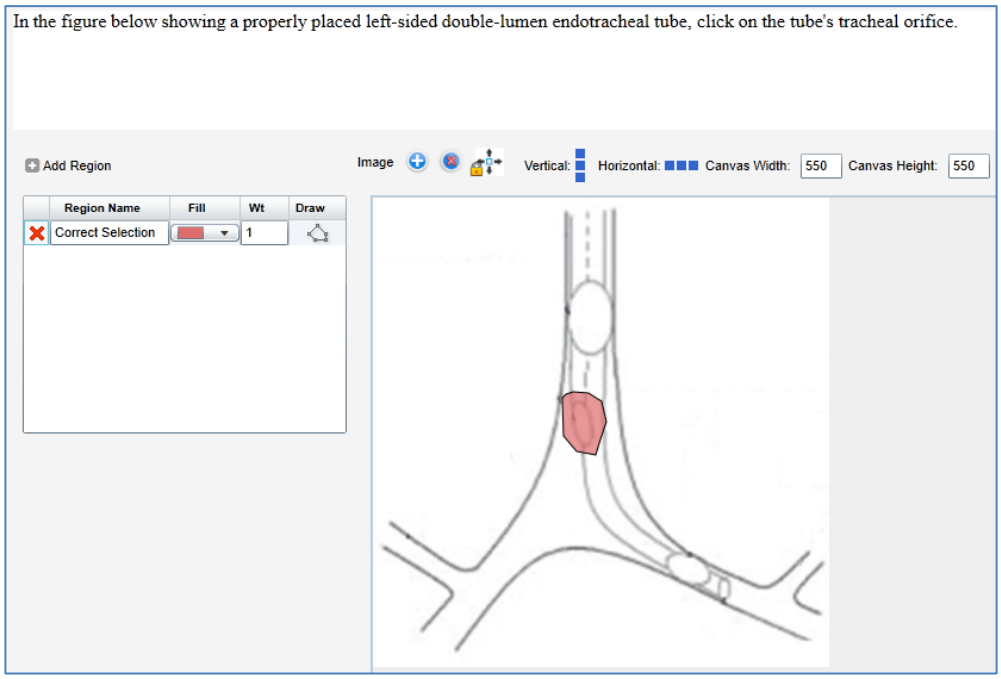
➡**Incorporate critical thinking wherever possible.** With hotspot questions, it is easy to get into the rut of basic anatomy/identification questions. For example, instead of:

In this diagram of the heart, click on the SA node…

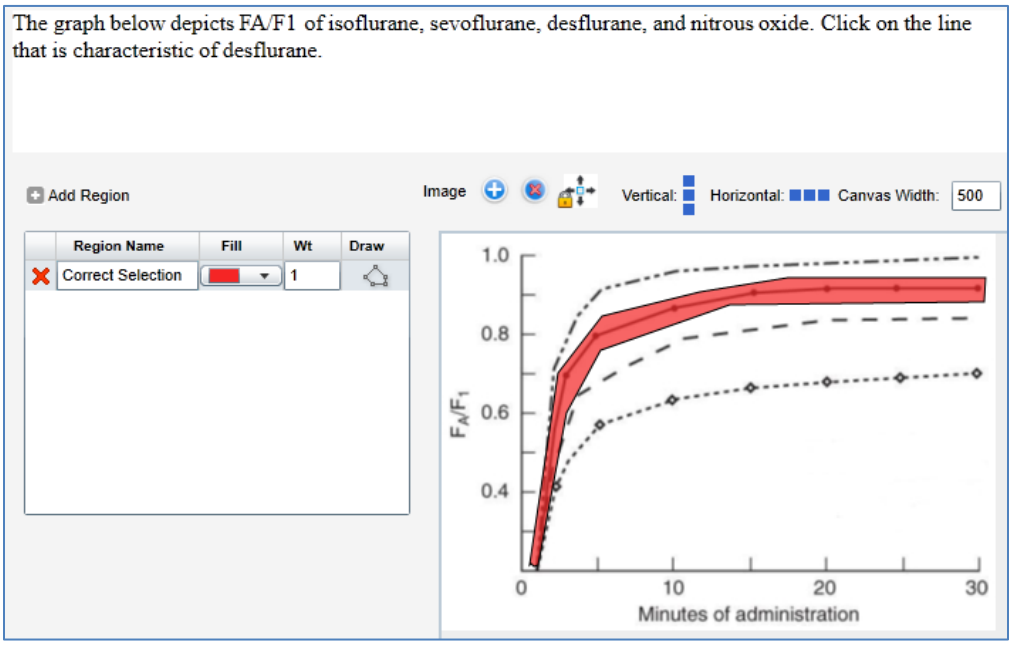
Try this:

In this diagram of the heart, click on the part of the heart that generates the sinus rhythm.

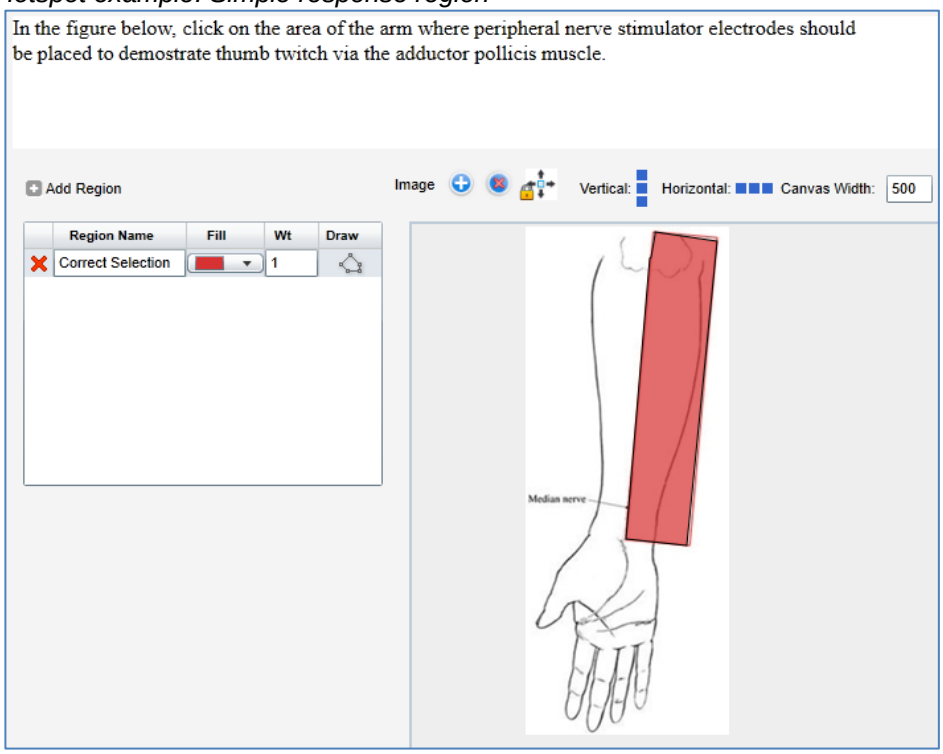
➡**Once the graphic is inserted, specify the correct region.**



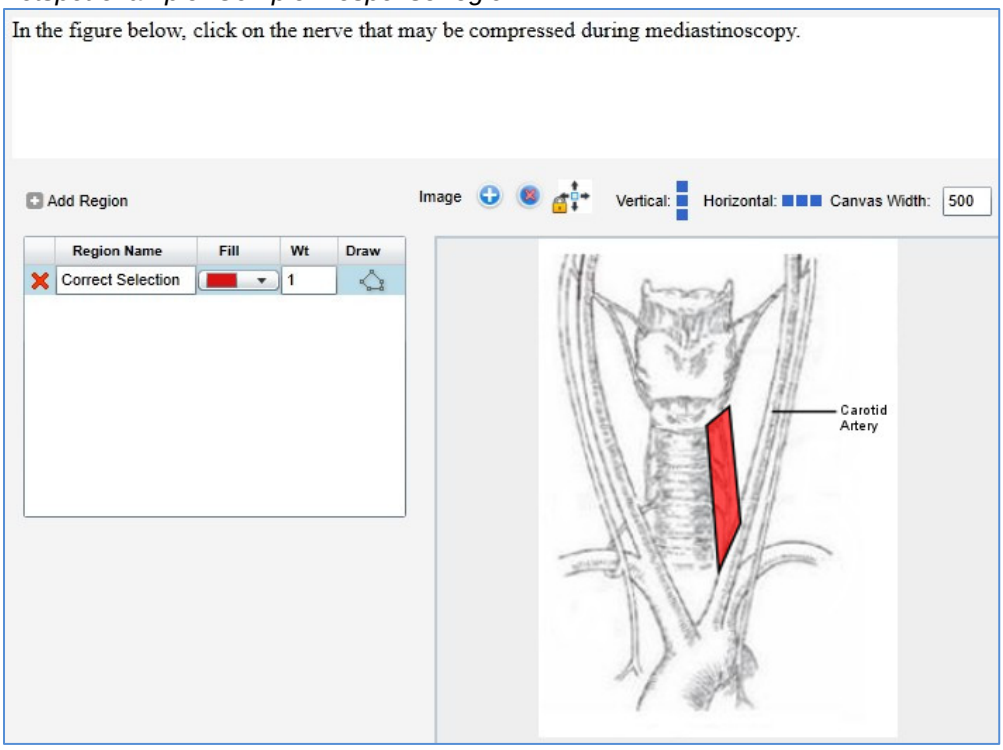
➡**Regions can be complex**—they do not have to be simple rectangles or ovals, etc.



Hotspot example: Simple response region

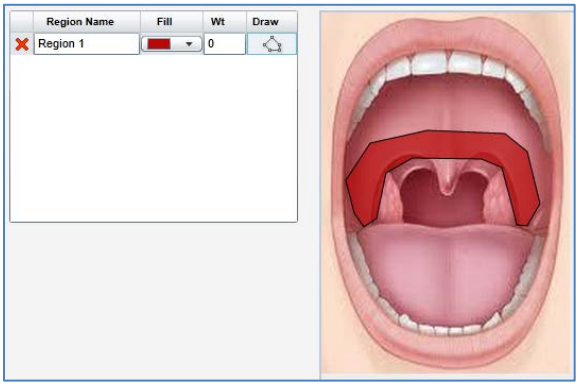


Hotspot example: Complex response region



➡**Ensure that a keyed area encompasses the entire range of correct responses without including incorrect areas**. In the example below, if the question asked for identification of the faucial pillars, the region marked for correct selection includes incorrect areas (i.e., pillars AND soft palate).

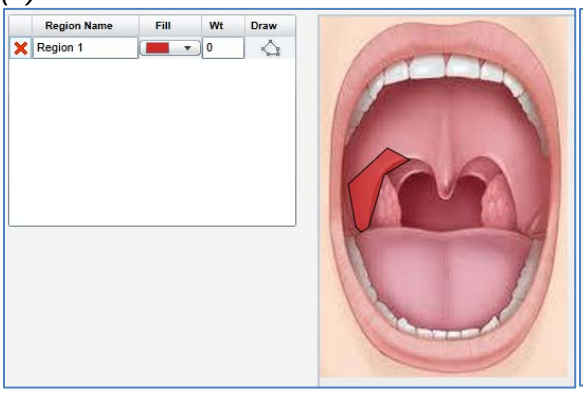
Hotspot example: Incorrect responses captured



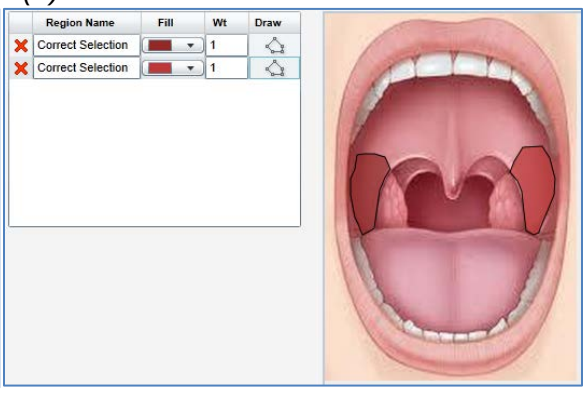
➡**Ensure that all correct responses are keyed areas.** Sometimes 1 region will not allow for coverage of all correct responses without the inclusion of incorrect responses. Using the example of a question that asked for the identification of the faucial pillars, 2 correct regions **must** be specified.

Hotspot example: (A) Correct responses not captured; (B) Correct responses captured

(A)



(B)



Developing Clinical Scenario Sets

Definition

Clinical scenario sets can be constructed in a variety of ways, but a prototypical clinical scenario set consists of a short "set header" paragraph describing a clinical situation, followed by a series of questions building on the initial scenario. Each question tests 1 additional decision or action and may add additional clinical information as the case progresses.

Reasons to Use Clinical Scenario Sets

Clinical scenario sets are generally used to try to measure constructs beyond simple knowledge, such as clinical decision-making ability. They should be well suited to access higher-level cognitive processes in Bloom's taxonomy, such as applying, analyzing, and evaluating. They present the opportunity to assess integrated knowledge as opposed to disparate points on a content outline. They are often used to enhance the face validity of an assessment as a measure of clinical skill.

Set Header Length.

A very short set header is unlikely to present sufficient detail to permit the candidate to understand the situation and answer questions requiring substantial analysis. The subsequent items would need to present the missing details, which can create issues as questions are added or removed from the set through revision and review.

Appropriately detailed set headers generally include:
- any relevant patient characteristics, such as age or gender
- presentation/context of the encounter
- relevant past history (comorbidities, medications, etc)
- physical examination findings or diagnostic results
- vital signs

These details permit the candidate to focus on particular problems that could be presented in the items that follow the set header.

A very long set header is likely to present different problems to the candidate. One major problem is the time required to read and process the set header. A candidate will use on average less than 1 minute responding to a test question. A long set header not only substantially increases examination time, but also permits the reading ability of the candidate to influence the test score. When reading ability substantially influences the exam score, the validity of the score is reduced.

Another problem with very long set headers is the amount of detail, both necessary and unnecessary, that the scenario presents. If too many details appear in the set header, the candidate will probably be unable to completely process the information required to answer the questions. In this situation, the test score becomes a composite of candidate knowledge and intelligence. Although intelligence is necessary to function in most professions, the certification exam is not the place to measure that skill.

To avoid undue set header length and minimize construct-irrelevant variance from reading, a tabular layout of patient data may be used.

Example:

> *An 85-year-old woman presents to the ambulatory surgery center for repair of a large ventral hernia under general anesthesia. She has hypertension and diabetes mellitus. Recent medications include*

lisinopril at bedtime and metoprolol this morning. She has not taken her glyburide today. The patient has been NPO per guidelines.

Vital signs
Blood pressure: 120/80 mm Hg
Heart rate: 60/min
Respiration rate: 16/min
SpO2: 96% on room air

Laboratory results [only include if relevant to case or important as distractor information]
Hemoglobin: 14.5 g/dL
Hematocrit: 41.7%
Platelets: 215,000/µL
Glucose, serum: 91 mg/dL
Blood urea nitrogen: 11 mg/dL
Creatinine, serum: 1.1 mg/dL
Albumin, serum: 4.8 g/dL
Bilirubin, total: 0.4 mg/dL
Alkaline phosphatase: 90 IU/L
Aspartate aminotransferase: 50 IU/L
Alanine aminotransferase: 65 IU/L

➡**A reasonable set header will usually have a length of 4 to 6 sentences.**

➡Writers should certainly try to use **graphics in set headers**, but just as when graphics are used in items, they **must be integral to the case, must be copyrightable by NBCRNA or capable of being used by permission, and must respect privacy rights of any patients** who are involved.

Developing Multiple-Choice Questions Linked to the Set Header.

➡**Scenario items must require the candidate to read the set header.** We achieve that link by making direct references to the set header in the item by naming (1) common problems and situations, (2) processes, and (3) interventions. Here is an example set header presenting reasonable detail to the candidate:

A patient diagnosed with carpal tunnel syndrome has carpal tunnel release surgery. Axillary brachial plexus block by transarterial approach is administered for regional anesthesia during the procedure, which is successful. Seven weeks following surgery, the patient has paresthesias in her hand and no recovery of motor function.

Following is an example of a linked question. The information required to correctly respond to the item appears explicitly in the set header.

Which of the following is the most appropriate initial diagnostic test regarding the patient's paresthesias?

A. X-ray
B. Ultrasonography
C. MRI
D. Nerve conduction study

Below is an example of an unlinked item. The candidate does not need to read the set header to answer this item, and asking the candidate such a question wastes the candidate's time taken to read the set header.

What is the reported incidence of brachial plexus injury in carpal tunnel release surgery?
A. 6%
B. 12%
C. 18%
D. 24%

➡**Questions that refer to the basic outline of the set header but involve information not found there are speculative.** Below is an example of a generally linked, but speculative item.

Which of the following technical factors in the patient's anesthesia is most likely responsible for her deficits?
A. Large-gauge needle trauma
B. Intraneural anesthetic injection
C. Axillary arterial laceration
D. Inappropriate arm positioning

If it is important for the candidate to think about circumstances of the scenario, information about those circumstances should appear with sufficient detail in the set header so the candidate can process the details of the scenario before engaging the items. The example, however, refers to information absent from the set header such as the gauge of needle used or the positioning of the patient's arm. This leaves the examinee with no grounds for a decision.

➡**It is preferred to have 4 to 8 questions per scenario.** One item per scenario defeats the purpose of using scenario-based items, which is to simulate a clinical experience and elicit the key actions and decisions in a situation. At the other extreme, using large numbers of items with a scenario presents 2 problems. First, if a scenario contains too many items all of the same classification, we could populate a test domain with that single scenario, impairing the balance of subtopics on the entire exam. Secondly, research on key feature problem writing suggests that only a few questions really assess clinical judgment in each scenario, so that extra questions are less likely to add information about the examinee's ability.

➡**Items for a given scenario do not have to have the same classification.** Usually overall test content can be balanced by modifying numbers of standalone items within the same test.

➡**Items for a scenario should not cue each other or give away answers to prior questions.**

Item Sensitivity Review

It is extremely important that all examinees have the same experience when testing so that an exam can be scored in a fair, consistent, and reliable manner. Bias can be introduced into test items inadvertently, however, and results in different exam performance levels among individuals who have the same ability.

Common sources of bias include the presence of terminology, content, or structure that is differentially familiar to individuals of particular genders, cultures, ethnicities, or religions.

To help detect and avoid bias, each item undergoes multiple stages of review and field-testing, ensuring that content clarity and comprehension is standardized among all test takers. The first line of defense for eliminating item bias is the item author. As items are developed, make certain to consider the following questions:

Content

- Does the item use content that could be unfamiliar to some groups?
- Will members of some groups get the item correct or incorrect for the wrong reason?
- Does the item use information or skill that might not be available within the required educational background of all examinees?
- Does the item use content that some people might have more opportunity to learn?

Representation

- Does the item give a positive representation of any race, ethnicity, religion, profession, gender, etc., that is mentioned?
- Does the item use controversial or inflammatory content?
- Does the item use demeaning or offensive content?
- Does the item use stereotypical occupations or stereotypical situations?

Language

- Does the item use unnecessarily difficult vocabulary and constructions?
- Does the item use group-specific language, vocabulary, or reference pronouns?
- Does the item present clues that might facilitate the performance of one group over that of another?

Thinking About the Test Takers

- Keep the vocabulary consistent with the test takers' levels of understanding.
- Minimize reading time in the phrasing of each item.
- Avoid trick items that mislead or deceive test takers into answering incorrectly.
- Avoid obscure concepts.
- Think about the characteristics that may limit or distort comprehension of the item, such as age, educational level, experiences, language, or physical or mental limitations.

Appendix: Overview of Statistical Analysis

Introduction

The material that follows includes a discussion of statistics used for evaluating items, both classical and item response theory (IRT) statistics, specific examples of statistics observed on actual examination questions, and how they would be interpreted.

Classical Statistics

Classical statistics are *population dependent*. This means that they apply only to the group of examinees or items that were used to generate them. They are very useful in describing the results of a particular test and for diagnosing the individual characteristics of an item. The most commonly used classical statistics are:

N—Number of Item Exposures

The number of exposures or "hits" is simply the number of times an item is selected from the pool of active items and administered to an examinee. Generally, a minimum of 100 hits is required to reliably calibrate an experimental item. Once an item is made active, a sample size of 30 is generally considered sufficient for continued use.

P-Value—Proportion Correct

Referred to as the *easiness* or *facility* of the item, this statistic is the percentage of examinees that selected a particular response. P-values provide the following general guidance:

If the P-value is low (in the range of 0.00 to 0.30), the item is very hard and should be reviewed to determine that it has not been miskeyed or that there is not more than one correct answer. If the P-value is greater than 0.95, the item is very easy. The item should be deleted or revised to present a greater challenge to the examinees.

A P-value can also be calculated for the response options (the correct answer and each of the distractors) by dividing the number of individuals that selected the response by the total number of examinees. If the P-value is approximately equal across all responses, examinees may be guessing at the answer. There may be insufficient information in the question to adequately determine the correct answer, or the content of the question may be too esoteric. If the P-value is zero for any response, this is called a *null distractor*. Null distractors are indicative of obvious answers, nonparallel distractors, or nonsensical distractors.

Point Biserial Correlation

Often referred to as *item discrimination*, the point biserial correlation is an association measure between an item result and the total test score. It is a correlation between the examinees who selected a particular response and the overall distribution of examinees across the test.

The interpretation of the point biserial provides general guidance in evaluating item performance. A positive point biserial is desired for the correct answer and a negative (or weakly positive) point biserial is desired for each incorrect response. Guidelines in the interpretation of the point biserial are as follows:

➡The point biserial for the correct response should be 0.10 or higher.

➡When the point biserial for the correct response is near 0.00, approximately equal numbers of more-able and less-able individuals selected that response.

➡A negative point biserial for the correct response and a positive point biserial for an incorrect response indicate that more-able examinees are selecting the incorrect response. The item may have been miskeyed, worded in an unclear manner, or obscure in content, or there may be more than one correct answer.

The point biserial tends to be less informative at the score extremes. This means that items having very low or very high P-values may have lower point biserials. Point biserial is also less useful with examinations having very small numbers of candidates. With a small test population, 1 or 2 more-able examinees selecting an incorrect response will cause the point biserials to change radically.

Median Response Time

Item response time, the number of seconds that elapse between a question's appearance on the screen and the examinee's submitted response, can be a very informative descriptor of item and examinee behavior. Inordinately long response time may suggest problems in item construction, while very short response time may suggest examinee foreknowledge of the item. Because the distribution of response times is not normal, the median of the distribution may serve as more stable index of central tendency than the mean.

Item Response Theory Statistics

Item Response Theory (IRT) provides a useful mathematical framework for the calibration and equating of items onto a common scale and the development of calibrated item banks. The explanation that follows presents information on the Rasch model or 1-parameter IRT model. Rasch model statistics are considered to be *population independent*.

With the Rasch model, more-able candidates have a greater probability of answering any question correctly than less-able candidates, and more-difficult items have a lower probability of being answered correctly than easy items. Ability estimates and difficulty calibrations calculated from an equation are placed onto a log-linear scale. The units of this scale are *logits* or log-odds-units. Rasch model statistics are:

Item Calibrations (Difficulty Calibrations)

Item calibrations represent the degree of difficulty of the item on the logit scale. This scale generally runs from about –2.5 to +2.5 for the NCE and is an equal-interval scale. Items with a negative difficulty are relatively easy, whereas items with positive difficulty are relatively hard. In this model, items retain their relative difficulty on the equal-interval scale regardless of the test sample used for analysis. This means that once different test administrations have been equated, item difficulty remains relatively constant and directly comparable across administrations.

Following are some guidelines for interpretation of item calibrations:

➡Estimates of item difficulty greater than +1.5 logits refer to hard items. These items should be reviewed to determine that they have not been miskeyed and do not have more than one correct answer.

➡Estimates of item difficulty less than –1.5 logits refer to easy items. They may have obvious answers or poorly written distractors, or be cued by information in other questions.

Displacement

In general, item difficulty calibrations remain constant over the course of an item's life, but fluctuations are possible, and it is important to periodically assess the stability of the estimates. The displacement of an item is the difference between the item's anchored difficulty and the calculated difficulty based on the data under analysis. Positive values indicate that the item got harder, while negative values indicate that the item got easier. Generally speaking, if an item calibration displaces more than ±0.75 logit, it is flagged for review.

Rasch Fit Statistics

The data-to-model fit statistics check how closely the performance of an item approximates the expectations of the Rasch model that more-able candidates will perform better on the item than less-able candidates and that harder items will be answered correctly less frequently than easier items.

The fit statistics are mean-squared residuals calculated as the ratio between the observed value and the value predicted by the model. The expected value is 1. Values greater than 1.3 indicate more than 30% unexplained noise in the item calibration. Guidelines for the interpretation of the fit statistics are as follows:

➡High mean-squared residuals (fit >1.3) indicate that the item has poor discrimination. The item may be worded ambiguously or have multiple correct answers.

➡Low mean-squared residuals (fit <0.70) indicate that the item is highly discriminating. These items should be reviewed, however, to make sure that they (1) do not contain especially seductive distractors, (2) do not require special knowledge to answer correctly, (3) do not involve a common misconception, and (4) are not cued by another item.

Overall Review of Item Statistics

Recommendations for item statistical criteria for the NCE and SEE are contained in Table 1 below. Note that recommendations for the 2 tests only differ with regard to difficulty.

Table 1

Statistic	Definition	Acceptable Range
P-value	The proportion correct	0.25 to 0.95 (Experimental) 0.60 to 0.80 (Active)
Point Biserial	The item response to raw score correlation	≥0.1
Diff (Difficulty)	The Rasch item difficulty	NCE: –2.5 (easy) to +2.5 (hard) SEE: –1.5 (easy) to +1.5 (hard)
Disp (Displacement)	The change in item difficulty	–0.75 (getting easier) to +0.75 (getting harder)
Count (N)	The number of people taking the item	>100 (Experimental) >30 (Active)
Fit (Infit and Outfit)	How well the item fits the model	<1.3
Std Err (Standard Error)	The precision of the item difficulty estimate	<0.3

Examples of Statistical Interpretation of Experimental Items

Example #1

The following item-level statistics were observed for an experimental item:

N_Total	Difficulty	Pval	PtBIS	Median_Time	MS_Infit
120.000	1.450	0.725	0.270	74.500	0.990

- The sample size was 120, so there were an adequate number of exposures to calibrate the item.
- The difficulty calibration is 1.45 logits, right in the middle of the difficulty range.
- Correspondingly, the P-value is 0.725, indicating about 73% of examinees responded correctly, so this item is moderately easy.
- The point biserial is 0.27, well above the 0.1 threshold. This item discriminates well between examinees of high and low ability.
- The median response time is about 1 minute and 15 seconds. This is a bit above the average response time on the NCE (38 seconds), indicating that examinees take a little longer to respond to the question.
- The infit statistic is very close to 1, meaning the item fits the expectations of the model.

Let's take a look at the response-level statistics for this item. For this item, the key is C:

#	Pval	PtBIS	N_Sel
A	0.080	–0.310	9.000
B	0.010	–0.040	1.000
C*	0.725	0.270	87.000
D	0.190	–0.090	23.000

- First, note that the P-value and point biserial for the key (C) are identical to the item-level P-value and point biserial.
- We see that each response option was selected at least once.
- The key has a positive PtBIS, and the distractors (A,B,D) have negative PtBIS.

Example 1 shows ideal response behavior.

Example #2

The following item-level statistics were observed for an experimental item:

N_Total	Difficulty	Pval	PtBIS	Median_Ti	MS_Infit
148.000	4.160	0.169	–0.060	22.000	1.10

- The sample size was 148, so there were an adequate number of exposures to calibrate the item.
- The difficulty calibration is 4.16 logits, very high on the difficulty scale. Correspondingly, the P-value is 0.169, indicating about 17% of examinees responded correctly, so this item is of high difficulty.
- The point biserial is –0.06, well below the 0.1 threshold. This item does not discriminate well between examinees of high and low ability. A negative biserial indicates that low-ability examinees answer the question correctly more often than high-ability examinees.
- The median response time is about 22 seconds, which is less than average (38 seconds). Given the difficulty of the question, the examinees may be guessing randomly.
- The infit statistic is very close to 1, meaning the item fits the expectations of the model.

Based on the high difficulty and poor discrimination of this item, it may not be appropriate to include as an active item.

Let's take a look at the response-level statistics for this item. For this item, the key is C:

#	Pval	PtBIS	Avg_Meas
A	0.000	—	—
B	0.740	0.260	2.550
C*	0.170	–0.060	2.380
D	0.090	–0.310	1.910

- The key has a negative PtBIS (–0.06), and only 17% selected it.
- Most people (74%) chose B, an incorrect answer. In fact B has the highest P-value and a positive PtBIS. This item may have been miskeyed.

Example 2 DOES NOT show ideal response behavior.

Examples of Statistical Interpretation of Active Items

For active (scored items), we evaluate different statistics for item performance. Because items have previously demonstrated good statistics, we are primarily concerned with the (a) the usefulness of the item, judging by sample size, and (b) stability of the difficulty calibration using the item displacement statistic. Generally, displacements greater than 0.5 logits are of concern.

Example #1
The following item-level statistics were observed for an active item:

N_Total	Difficulty	Disp	Pval	PtBIS
264.000	1.305	0.180	0.674	0.140

- The sample size was 264, so there were an adequate number of exposures to assess the statistics. The item is being seen by a good sample of people.
- The original difficulty calibration is 1.305 logits, in the middle-to-high end of the difficulty scale.
- Correspondingly, the P-value is 0.674, very close to the expected value of 70%.
- The displacement statistic is 0.18. This is a small displacement value, and insignificant. The item difficulty is stable.

Example #2

N_Total	Difficulty	Disp	Pval
60.000	2.550	–0.790	0.820

- The sample size was 60, so there were an adequate number of exposures to assess the statistics. The item is being seen by a good sample of people.
- The original difficulty calibration is 2.55 logits, moderately difficult.
- The displacement statistic is –0.79. This is a large displacement value, and significant. For some reason, this item has become substantially easier. Correspondingly, the P-value is 0.82, somewhat higher than the expected value of 70%.

Final Word on Statistical Review of Items

Item statistics are reviewed after each examination or in the case of an adaptive test, on a periodic basis. Any item that exhibits unusual statistical performance is flagged for further review by content experts. Statistics are only a guide to item performance. Content experts must make the final determination as to whether items will be retained, deleted from scoring, reworked, or retired from the item bank.