Chapter 2
Multiple-Choice-Item Formats

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

True/False vs One-Best-Answer Questions

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

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

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

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

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

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

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

The True/False Family

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

The following is an example of an acceptable true/false item from a structural perspective.* Note that the stem is clear and the options are absolutely true or false with no ambiguity.

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

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

The options can be diagramed as follows.

2
4

Totally Wrong Options

1
3

Totally Correct Options

True statements about cystic fibrosis (CF) include:

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

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

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

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

The way to a man's heart is through his

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

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

In the clinical assessment of chronic pain,

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

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

In children, ventricular septal defects are associated with

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

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

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

- Stems must be clear and unambiguous. Imprecise phrases such as *is associated with*; *is useful for*; *is important* and words that provide cueing such as *may* or *could be*; and vague terms such as *usually* or *frequently* should be avoided.

- Options must be absolutely true or false; no shades of gray are permissible; avoid phrases and words noted in the first item above.

The One-Best-Answer Family

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

Stem:

A 32-year-old man has a 4-day history of progressive weakness in his extremities. He has been healthy except for an upper respiratory tract infection 10 days ago. His temperature is 37.8 C (100 F), blood pressure is 130/80 mm Hg, pulse is 94/min, and respirations are 42/min and shallow. He has symmetric weakness of both sides of the face and the proximal and distal muscles of the extremities. Sensation is intact. No deep tendon reflexes can be elicited; the plantar responses are flexor.

Lead-in:

Which of the following is the most likely diagnosis?

Options:

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

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

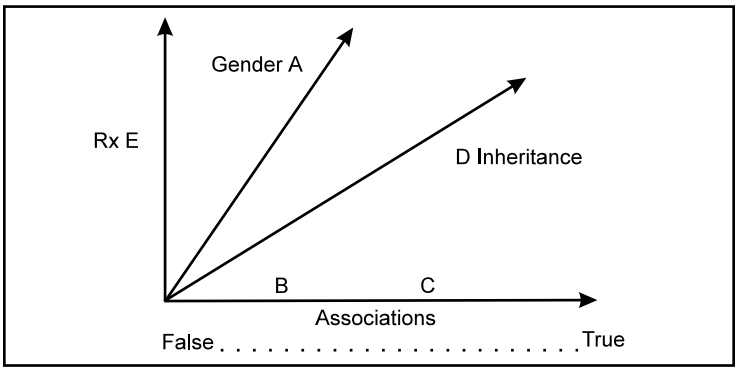
D	C	A	E	B
Least Correct				***Most Correct***

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

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

Which of the following is true about pseudogout?

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



The diagram of these options might look like this. The options are heterogeneous and deal with miscellaneous facts; they cannot be rank-ordered from least to most true along a single dimension. Although this question appears to assess knowledge of several different points, its inherent flaws preclude this. The question by itself is not clear; the item cannot be answered without looking at the options.

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

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

The Bottom Line on Item Formats

We recommend that you do not use true/false questions. While many item writers believe the true/false items are easier to write than one-best-answer items, we find that they are more problematic. The item writer had something particular in mind when the question was written, but careful review commonly reveals subtle difficulties that were not apparent to the item author. Often the distinction between "true" and "false" is not clear, and it is not uncommon for subsequent reviewers to alter the answer key. As a result, reviewers rewrite or discard true/false items far more frequently than items written in other formats. Some ambiguities can be clarified, but others cannot.

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

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

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

Chapter 3
Technical Item Flaws

❖

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

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

Issues Related to Testwiseness

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

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

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

A. examination of cerebrospinal fluid
B. glucose with vitamin B_1 (thiamine)
C. CT scan of the head
D. phenytoin
E. diazepam

Logical cues: a subset of the options are collectively exhaustive

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

Crime is

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

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

In this item, Options A, B, and E contain terms that are less absolute than those in Options C and D. The testwise student will eliminate Options C and D as possibilities because they are less likely to be true than something stated less absolutely. Note that this flaw would not arise if the stem was focused and the options were short; it arises only when verbs are included in the options rather than in the lead-in.

In patients with advanced dementia, Alzheimer's type, the memory defect

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

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

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

Secondary gain is

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

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

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

A 58-year-old man with a history of heavy alcohol use and previous psychiatric hospitalization is confused and agitated. He speaks of experiencing the world as unreal. This symptom is called

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

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

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

Local anesthetics are most effective in the

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

Issues Related to Irrelevant Difficulty

Options are long, complicated, or double

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

not the student can interpret and synthesize information to determine, for example, the most likely diagnosis, then it is appropriate for the stem to include a fairly complete description of the situation.

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

A. Notice, an impartial forum, council, a chance to hear and confront evidence against him/her.

B. Proper notice, a tribunal empowered to make the decision, a chance to confront witnesses against him/her, and a chance to present evidence in defense.

C. Reasonable and timely notice, impartial panel empowered to make a decision, a chance to hear evidence against himself/herself and to confront witnesses, and the ability to present evidence in defense.

Numeric data are not stated consistently

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

In this example, Options A, B, and C are expressed as ranges, whereas Options D and E are specific percentages. All options should be expressed as ranges or as specific percentages; mixing them is ill-advised. In addition, the range for Option C includes Options D and E, which almost certainly rules out Options D and E as correct answers.

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

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

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

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

Severe obesity in early adolescence

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

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

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

In a vaccine trial, 200 2-year-old boys were given a vaccine against a certain disease and then monitored for five years for occurrence of the disease. Of this group, 85% never contracted the disease. Which of the following statements concerning these results is correct?

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

"None of the above" is used as an option

The phrase "None of the above" is problematic in items where judgement is involved and where the options are not absolutely true or false. If the correct response is intended to be one of the other listed options, knowledgeable students can be faced with a dilemma because they have to decide between a very detailed perfect option and the one that you have intended as correct. They can often construct an option that is more correct than the one you intended to be correct. Use of "none of the above" essentially turns the item into a true/false item; each option has to be evaluated as more or less true than the universe of unlisted options. It will often be possible to fix such items by replacing "none of the above" by an option that means roughly the same thing but is more specific. For example, in an item asking an examinee to specify the most appropriate pharmacotherapy, replacing "none of the above" by "no drug should be given at this time" will eliminate the ambiguity of "none of the above."

Which city is closest to New York City?

A. Boston

B. Chicago

C. Dallas

D. Los Angeles

E. none of the above

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

Stems are tricky or unnecessarily complicated

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

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

I: 46, XX, -14, +T (14q21q) pat
II: 46, XX, -14, +T (14q21q) de novo
III: 46, XX, -14, +T (14q21q) mat
IV: 46, XX, -21, +T (14q21q) pat
V: 47, XX, -21, +T (21q21q) (parents not karyotyped)

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

Summary of Technical Item Flaws

Issues Related to Testwiseness

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

Issues Related to Irrelevant Difficulty

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

General Guidelines for Item Construction

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

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

Use of Imprecise Terms in Examination Questions

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

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

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

True statements about pseudogout include:

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

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

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

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

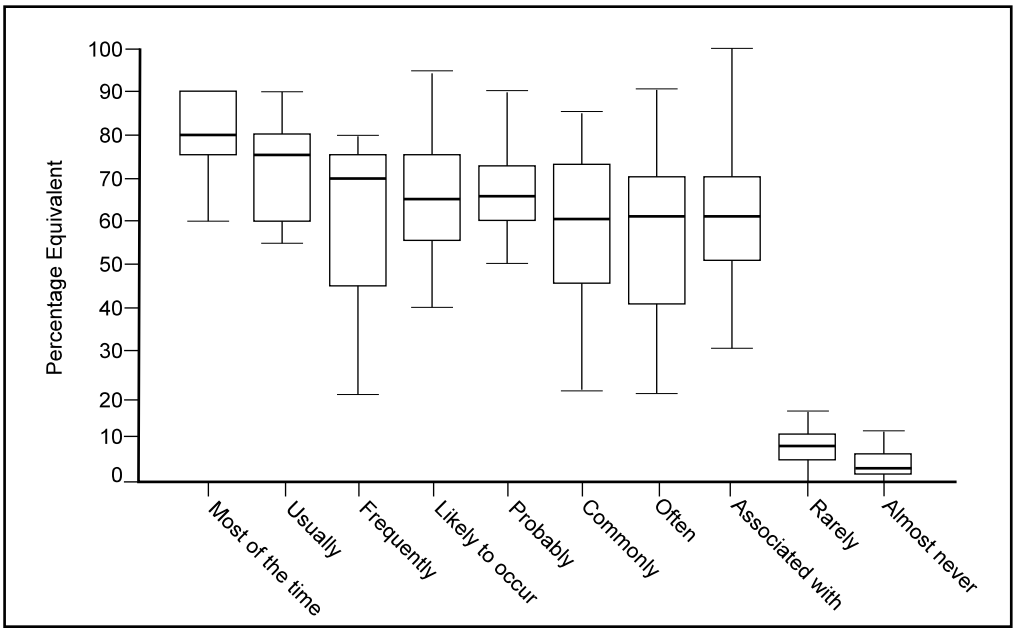
Which of the following laboratory values is usually increased in patients with pseudogout?

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

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

Patients with pseudogout have pain:

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



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

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

The Basic Rules for One-Best-Answer Items

- ***Each item should focus on an important concept, typically a common or potentially catastrophic clinical problem.*** Don't waste testing time with questions assessing knowledge of trivial facts. Focus on problems that would be encountered in real life. Avoid trivial, "tricky," or overly complex questions.

- ***Each item should assess application of knowledge, not recall of an isolated fact.*** The item stems may be relatively long; the options should be short. Clinical vignettes provide a good basis for a question. For the clinical sciences, each should begin with the presenting problem of a patient, followed by the history (including duration of signs and symptoms), physical findings, results of diagnostic studies, initial treatment, subsequent findings, etc. Vignettes may include only a subset of this information, but the information should be provided in this specified order. For the basic sciences, patient vignettes may be very brief; "laboratory vignettes" are also appropriate.

- ***The stem of the item must pose a clear question, and it should be possible to arrive at an answer with the options covered.*** To determine if the question is focused, cover up the options and see if the question is clear and if the examinees can pose an answer based only on the stem. Rewrite the stem and/or options if they could not.

- ***All distractors (ie, incorrect options) should be homogeneous.*** They should fall into the same category as the correct answer (eg, all diagnoses, tests, treatments, prognoses, disposition alternatives). Rewrite any dissimilar distractors. Avoid using "double options" (eg, do W and X; do Y because of Z) unless the correct answer and all distractors are double options. Rewrite double options to focus on a single point. All distractors should be plausible, grammatically consistent, logically compatible, and of the same (relative) length as the correct answer. Order the options in logical order (eg, numeric), or in alphabetical order.

- ***Avoid technical item flaws that provide special benefit to testwise examinees or that pose irrelevant difficulty.***

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

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