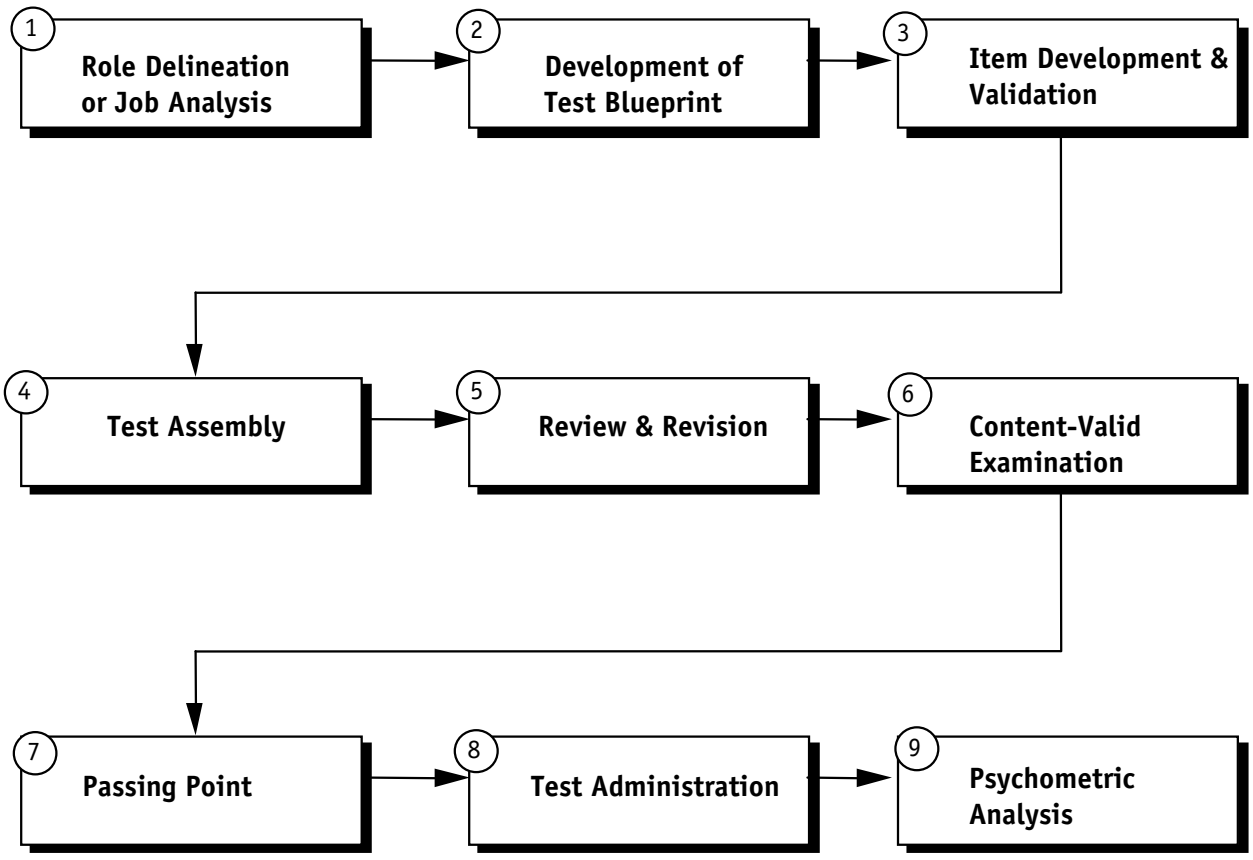
Technical Guidelines for the Construction of Multiple-Choice Questions

Including
Developing a Variety of Multiple-Choice Items

Contents

Steps in the Construction of a Content-Valid Examination



Overview of the Test Development Process

Role Delineation. Before developing an examination, a role delineation study determines the knowledge and skills that define a minimally competent professional in the field to be tested. Linking the knowledge and skills defined in the role delineation study to the examination content ensures that an examination is content valid. In psychometric terms, validation is how a test developer documents the competence inferred from an examination test score.

During the role delineation process, a committee of subject matter experts defines the overall performance domains associated with competent practice. These performance domains are further broken down into more distinct tasks, knowledge, and skills required on the job. The job responsibilities developed by the subject matter experts are then validated through a survey of practitioners. The practitioners review and rate the domains and tasks according to their importance, criticality, and frequency of performance.

Development of Test Blueprint. In the next step, the results from the validation survey are used to develop a blueprint, or a plan, for the examination. The information regarding the importance, criticality, and relevance of each domain and task is translated directly into the percentage of items that should be included in the examination for each content area. 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.

Item Development and Validation. All examination items are written by experts in the practice field. Each item writer is trained in writing, reviewing, editing, and validating questions. Each question is reviewed and validated by at least three other subject matter experts and must have a verifiable reference. Each item is classified by content category, assigned a cognitive level, and validated according to its appropriateness to the certification-level practitioner. After development, items are reviewed to ensure they are psychometrically sound and grammatically correct.

Test Assembly. Each examination is created by selecting the appropriate number of items from each content area, as specified in the test blueprint.

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

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.

Passing Point. A valid credentialing examination must have a defensible passing score. The cut-off score that separates examinees who pass from examinees who fail must be based on the minimum competence required to protect the public from harm. A criterion-referenced approach called the Modified Angoff Technique is often used to determine the cut score or passing point of an examination. This technique is currently considered by the testing profession to be one of the most defensible criterion-referenced methods available for setting passing points.

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, rest room facilities, handicap accessibility, and noise control.

Psychometric Analysis. Following the test administration, the item statistics are reviewed to ensure quality and validity. Item 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.

Item Writing Guidelines

1. The stem should be meaningful in and of itself and present a definite problem. A poor stem is:

The Middle Ages:

A. Saw the rise and fall of the Roman Empire and many literary and cultural changes.
B. Is the period of European history that lasted from about 350 A.D. to 1450 A.D.
C. Ended drastically with the beginning of the Renaissance period of European history.
D. Was when universities and cities first began to take their current modern form.

A better stem is:

The period of European history that lasts from about 350 A.D. to 1450 A.D. is known as the:

A. Middle Ages.
B. Black Death.
C. Renaissance.
D. Reformation.

2. The stem should include as much of the item as possible and be free of irrelevant material. The intent of the stem is to focus the test-taker directly on the tested information. A poor stem is:

The Middle Ages was followed by the Renaissance period of European history. What marked the change between them?

A better stem is:

What marked the change between the Middle Ages and the Renaissance period of European history?

An even better stem for this example is:

The change between the Middle Ages and the Renaissance period of European history was marked by

3. Avoid negatively stated stems and words such as NOT or EXCEPT. Negatively stated items cause the candidate to take an extra cognitive step in the consideration of each response. This extra step generally renders negatively worded items more difficult than the corresponding positively worded item. However, that extra degree of difficulty is linked more to the ability of the candidate to perform mental manipulations than to the knowledge of the candidate in the area being tested.

A poor stem would be:

Which state is **NOT** located north of the Mason-Dixon line?

A better stem is:

Which state is located south of the Mason-Dixon line?

4. **All response options should be grammatically consistent with the stem. Candidates generally assume that the correct response will most likely be grammatically consistent, and if there are options that are inconsistent, test-wise candidates can eliminate them quickly.**

An item in which options do not all follow grammatically from the stem is:

An electric transformer can be used:

A. For storing electricity.
B. To increase the voltage of alternating current.
C. It converts electrical energy into mechanical energy.
D. Alternating current is changed into direct current.

A better example for this question is:

An electric transformer can be used to:

A. Store electricity.
B. Increase the voltage of alternating current.
C. Convert electrical energy into mechanical energy.
D. Change alternating current to direct current.

5. **All distractors must be plausible. In writing distractors, item writers may use common errors, important sounding words (significant, accurate, etc.), textbook-type language, and words with verbal associations in the stem (politician...political). Item writers may think of common misunderstandings and careless errors. Distractors must be drawn from the same pool, which means that they are all members of the same group as the correct response. Distractors must be similar in length, complexity, vocabulary, grammatical construction, etc. An item with distractors that are not drawn from the same pool as the correct response is:**

The state bird of North Carolina is the:

A. Robin.
B. Penguin.
C. Cardinal.
D. Purple Martin.

A better example is:

The state bird of North Carolina is the:

A. Robin.
B. Wild turkey.
C. Cardinal.
D. Purple Martin.

6. **Do not use "All of the above" and "None of the above" as distractors. If a candidate can eliminate one of the other three distractors, then "All of the above" is also eliminated. Similarly for "None of the above." In such an event, the chances of the candidate guessing the correct response are substantially improved.**

7. **Avoid using terms like "always" and "never." Very few things in professional settings are always true or never true. Candidates can use these terms in eliminating distractors.**

8. **When writing questions, include a published reference where the content of each question can be verified. If possible, include more than one published reference for each question. An examination containing one unreferenced item is legally indefensible.**

9. **Each question must be classified according to an appropriate classification number. The classification number represents the content area that a specific question is testing. Each classification number contains six digits. The first two digits represent the domain being tested, the second two digits represent the task statement being tested, and the last two digits represent the knowledge or skill statement being tested. For example, classification number 010203 would represent domain 1, task 2, knowledge and skill statement 3. Every item must reflect a single specific knowledge or skill, as required in the test specifications. The courts of the United States (and most western countries) 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 one item that is not classified is not legally defensible.**

10. **Avoid overly general content. Each item should have a specific purpose focused on a specific content area.**

11. **Avoid opinion-based questions. Items that require a candidate to exercise professional judgment are important, but such items should focus on well-established guidelines. Judgment using the work and opinion of obscure and controversial professionals should be avoided by item writers.**

12. **Avoid trick items. The intent of the examination is to assess the knowledge and skill of the candidate, not to determine how clever the candidate (or item writer) can be.**

13. **Use vocabulary appropriate for the target audience. If the candidate population generally has a high school diploma, then the use of vocabulary learned in graduate school is not appropriate. Use acronyms sparingly. When in doubt about an acronym, spell it out in parentheses.**

14. **Every question written must have the correct number of response options. All the examples in this booklet use four response options for each question. However, some certification programs require more 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.**

15. **Ensure that the directions in the stem are very clear. The intent of the stem is to focus the candidate on the decision that will demonstrate the knowledge of the candidate. Some "smoke" in stems that describe a situation can be appropriate because many professional situations can be unclear on first encounter. However, remember that the candidate does NOT have the option to consult peers and mentors before answering an examination question.**

16. **Avoid excessive verbiage. Item writing is not the occasion to demonstrate you prowess at writing like Faulkner.**

17. **Make sure that each question only has one correct option. Items that require the candidate to exercise professional judgment can present a problem here in that degrees of correctness are important. Item writers should consider common misunderstandings of best practice with items involving judgment. Regional differences in practice should be avoided as options.**

18. **Place options in logical or numerical order. If the responses are 1, 3, 6, and 9, list them that way, or the reverse. Order alphabetical responses the same way.**

19. **Keep options independent; choices should not be overlapping. A sufficient degree of independence can be difficult with items that require professional judgment, and review by multiple item writers might be necessary before the options are sufficiently independent for good item performance.**

20. **Keep the length of options about equal. A common mistake of item writers is to make the correct response far longer than the distractors.**

21. **Use typical errors to write distractor options. If an item involves the computation of a square root, then the incorrect answer reached by squaring might make a good distractor. Similarly, if an item requires a conversion between metric units, then a plausible distractor might be reached by shifting the decimal place.**

Types of Multiple-Choice Questions

There are three basic types of multiple-choice questions: direct questions, incomplete statements, and best answer items. These three types of questions can be used in simple multiple-choice as freestanding items, scenario items, interpretive exercises, and other formats. They can also be used to assess candidates' recall of factual knowledge, application, and analysis. Item writers should select which of the three types will make the item clearest to the candidate. *Do not use matching, true-false, or complex multiple-choice (K-type or Roman numeral) questions.*

In direct multiple-choice questions, the stem states a complete problem in the form of a question. The options, although plausible, contain one correct response and three definitely incorrect distractors. An example of a direct question is:

Which of the following cities is the capital of Florida?

A. Jacksonville
B. Tallahassee
C. Miami
D. Tampa

Direct questions generally include the phrase "which of the following" to direct the candidate to the consideration of the listed options, not to the consideration of the universe of possible answers. The "which of the following" construction can be omitted and replaced with "what" when the item requires precise thought governed by rules that do not involve judgment (e.g., what is 2 times 3?).

Other multiple-choice questions might be better written as incomplete sentences. The stem poses a definite problem, but allows the candidate to complete the sentence by selecting the correct option. As with direct questions, the options available for the candidate's selection include one correct response, with the distractors all being plausible but clearly incorrect. An example of an incomplete statement is:

The capital of Florida is:

A. Jacksonville.
B. Tallahassee.
C. Miami.
D. Tampa.

In cases where the candidate bases the selection of the response on the relative degree to which options are correct, item writers should use the best answer format. The correct response must be the best of the alternatives, whereas the distractors are correct to a definitely lesser extent. The stem may be a question or incomplete sentence. An example of the best answer format for multiple-choice questions is:

Which of the following factors contributed **MOST** to the selection of Tallahassee as the capital of Florida?

A. Location
B. Climate
C. Highways
D. Population

Developing Scenarios and Related Items

The length of the scenario is important. A very short paragraph is unlikely to present sufficient detail to permit the candidate to understand the situation the writer seeks to describe. For example, "You are a project manager for an international fertilizer company that plans to upgrade 32 servers distributed through 14 countries." Although this sentence describes a particular situation, a candidate will be unlikely to glean sufficient detail to be able to answer questions requiring substantial analysis, and the subsequent items would need to present the missing details, which defeats the intent behind developing scenario items.

This short scenario might be better if it were lengthened to include additional details regarding computer applications, hardware, personnel, time zones, deadlines, and languages. Such details as these would permit the candidate to focus on particular problems that could be presented in the items that follow the scenario.

A very long paragraph is likely to present different, though equally serious, problems to the candidate. Not the least of these problems is the time required to read and process the scenario. We generally anticipate a candidate will use on average less than one minute responding to a test question. A scenario requiring many minutes to process not only substantially increases the time required to finish the exam but also permits the reading ability of the candidate to influence the test score. Granted, a candidate has to be able to read to pass the exam; however, the burden of reading should not present a greater task than is presented by the content of the exam. In the event that candidate's reading ability substantially influences the exam score, the reliability of that score is necessarily reduced because of the two-dimensional nature of the score.

Another problem with very long scenarios is the amount of detail, both necessary and unnecessary, the scenario presents. If too many necessary details appear in the scenario, it is likely that the candidate will not be able to completely process the important associations required to answer the questions. In this situation, the test score becomes a two-dimensional composite of candidate knowledge and candidate intelligence, and although a degree of intelligence is necessary to function in most professions, the certification exam is not the place to measure candidate intelligence.

Sometimes item developers construct long scenarios containing too many unnecessary details in an effort to present a confusing situation similar to those often encountered in complex professional work. However, the inclusion of too many extraneous details requires the candidate to efficiently sort the important details from the unimportant, and this process permits candidate intelligence to become a part of the test score. In this instance and in the one in the last paragraph, the reliability of the test score will be reduced by the inclusion of the second unintended dimension of intelligence. This reduction in reliability will, in turn, reduce the validity of the inference we make from test scores about candidate competence.

There is little research to indicate the optimal length of the scenario used in testing. However, using brief sentences, a reasonable scenario will usually have a length of five or six sentences.

Developing multiple-choice questions linked to the scenarios. It is important that the test item requires the candidate to read the scenario. We achieve that link by making direct references to the scenario in the item. Some ways to make that reference are (1) the use of personal names, (2) the use of formal names of common problems and situations, (3) the use of the accepted names of processes, and (4) the use of product names. Here is an example scenario presenting reasonable detail to the candidate:

You manage a small catfishing project on the Neuse River. You are planning an excursion for this coming Saturday night, when you'll fish along a 100-yard stretch of the river. You plan to use a rope (trotline) tied over the water along this part of the river, and you will hang fishing lines, each one yard in length, along the length of this rope. Each line will have one hook. You will space the fishing lines about 2 yards apart on the rope to prevent adjacent lines from tangling with each other. You plan to use half of a chicken liver to bait each hook.

Following is an example of a linked question. The information required to correctly respond to the item appears explicitly in the item. Note that the use of "approximately" permits the candidate to make some minor assumptions about rope length lost to knots without putting the candidate at risk of responding incorrectly.

Approximately how many hooks will you place on the trotline?

A. 25
B. 50
C. 75
D. 100

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

What is 100 divided by 2?

A. 25
B. 50
C. 75
D. 100

Here is another example of a linked item. This item would not likely appear concurrently with the previous linked item because they both require the candidate to compute the number of hooks used on the trotline, which would result in two inappropriately correlated items. We say this correlation is inappropriate because the process to answer the item correctly overlaps. That is, both items involve computing the number of hooks. We need to limit the degree of overlap between items no more than the overlap produced by reading the scenario.

Approximately how many chicken livers will you need to bait each hook once?

A. 25
B. 50
C. 75
D. 100

Here is an example of an extraneous item. If it is important for the candidate to think about additional circumstances with the scenario, information about those circumstances should appear with sufficient detail in the paragraph, so the candidate can process the details of the scenario before engaging the items.

> How many hooks do you expect to lose to eels?
>
> A. 25
> B. 50
> C. 75
> D. 100

Here is an example of the previous scenario written poorly. Such a scenario would require additional information in the item stems, and would also require the candidates to infer too much detail, likely resulting in the candidate interpreting the scenario incorrectly for the intent of the subsequent items.

> *You are planning a catfishing trip for the next Saturday night. You plan to fish using a trotline along the bank of a river, using chicken livers for bait.*

In addition, we usually try to have three questions per item, though we can work with two or four items per scenario. One item per scenario begins to defeat the purpose of using scenario-based items where candidates need to exhibit analytical skills to answer the items.

Using more than four items with a scenario presents at least two problems. The first problem is technical, and involves the independence of test items. Item independence is superficially the case where one item does not contain a hint to another item. At a deeper level, scenario-based items are necessarily dependent because of the scenario. However, we are willing to accept a small degree of dependence as a sacrifice necessary to permit the efficient use of items requiring professional analysis.

The second problem is quite practical, and involves test construction in which a test blueprint specifies the proportions of items to use with each domain and task. If one scenario contained 200 items all of the same classification, we could populate the test with that single set, leaving no freedom to meet the requirements of the examination blueprint.

Writers often ask about item classification when writing scenario-based items, and whether all the items for a given scenario need to have the same classification. The strict answer is no. However, when all the related items fall into the same domain and task, the subsequent job of test construction is often easier, so we generally encourage item writers to keep the related items in the same domain and task, though we permit some exceptions by necessity.

Testing Within the Cognitive Domain

Benjamin Bloom and his associates developed taxonomies to describe various levels of cognitive, affective, and psychomotor activity. Because multiple-choice tests pertain most to the cognitive domain in which people store and use information to solve problems, item writers should write multiple choice questions that assess the candidate's ability at least at the first three levels (knowledge, application, and analysis) rather than simply at the knowledge level.

Assessing Recall/Understanding

Item writers can develop a variety of questions at the knowledge level. Professional knowledge is the body of terms, facts, and principles that characterize the profession. It is information that is stored in books, other media, and memory.

One type of recall/understanding question involves **terminology** because the knowledge base for a profession may be identified by its jargon. Even though candidates might not have large general vocabularies, their competent performance of job duties often requires that they possess the specific vocabulary of the field. An example of a multiple-choice question assessing recall or understanding of terminology is:

> Which of the following words has the same meaning as egress?
>
> A. Depress
> B. Enter
> C. Exit
> D. Regress

Professions are also distinguished by the body of **factual knowledge** that practitioners use in supplying competent service to their clients. The knowledge base is built through ongoing research, and candidates should demonstrate their mastery of factual knowledge. An example of a multiple-choice question assessing factual knowledge is:

> Which of the following missiles launched the first U.S. astronaut into orbital flight around the earth?
>
> A. Atlas
> B. Mars
> C. Midas
> D. Polaris

Normally, however, the recall of facts is insufficient for competent professionals. They must also possess **understanding of important principles.** Principles explain phenomena: how things work, how things move, why things are true, etc. An example of a multiple-choice question assessing knowledge of principles is:

The principle of capillary action helps to explain how fluids:

A. Enter solutions of lower concentration.
B. Escape through small openings.
C. Pass through semipermeable membranes.
D. Rise in fine tubes.

In addition, the competent professional must have knowledge of the **methods and procedures** by which professional services are delivered. As with terminology, facts, and principles, professions are often distinguished from each other by the methods commonly used. Knowledge of terminology, facts, and principles does not mean that candidates are knowledgeable about techniques for service delivery. An example of a multiple-choice question assessing knowledge of methods and procedures is:

If you were making a scientific study of a problem, your **first** step should be to:

A. Collect information about the problem.
B. Develop hypotheses to be tested.
C. Design the experiment to be conducted.
D. Select scientific equipment.

Assessing Application

Multiple-choice questions that assess application pose a problem or circumstance that professionals might reasonably experience in performing their responsibilities. The problem or circumstance presented in the question illustrates relevant facts or principles and the relationship between the factor or principle and some phenomenon. If the problem involves a method or procedure, application-level questions might ask the candidate to identify the correct justification for the method or procedure.

Many application-level questions require the candidate to recognize **the correct application of facts and principles.** These questions present a phenomenon, and then require the candidate to identify the fact or principle it represents. Other questions **present the fact or principle, and then require the candidate to choose from among several phenomena presented.** An example of a multiple-choice question assessing the candidate's identification of correct application is:

Which of the following statements illustrates the principle of capillarity?

A. Fluid is carried through the stems of plants.
B. Food is manufactured in the leaves of plants.
C. The leaves of deciduous plants lose their color in winter.
D. Plants give off moisture through their stomata.

Multiple-choice questions can require candidates to **identify the cause of some event** (the fact or principle explaining why or how the event occurred). They can also ask the candidate **to identify the event that will result from a cause.** An example of a multiple-choice question assessing the ability to interpret cause and effect relationships is:

Bread does **not** become moldy as rapidly if placed in a refrigerator because:

A. Darkness retards the growth of mold.
B. Cooling prevents the bread from drying out quickly.
C. Mold requires both heat and light for best growth.
D. Cooling retards the growth of fungus.

When the topic of a question is a method or procedure, the problem might be **to select the correct method for the problem** presented in the question or **to recognize why the method or procedure is appropriate in a given situation.** An example of a multiple-choice question assessing the justification of methods or procedures is:

Which of the following explains why adequate lighting is necessary in a balanced aquarium?

A. Fish need light to see their food.
B. Fish take in oxygen in the dark.
C. Plants expel carbon dioxide in the dark.
D. Plants grow too rapidly in the dark.

Assessing Analysis

Analysis refers to the ability to break information into its component parts so that its organizational structure can be understood. This can include the **identification of the parts, analysis of the relationship between the parts, and recognition of the organizational principles involved.** Candidates must demonstrate their understanding of both the content (terminology, facts, principles, and methods and procedures) and the structural form (relationships among facts, principles, and methods and procedures). In analysis, the candidate must recognize unstated assumptions and logical fallacies. Analysis-level questions require the candidate **to distinguish between facts and inferences, evaluate the relevance of certain material, and reach conclusions about the organizational structure of the material presented.**

When questions require candidates to review information and **distinguish fact from inference,** they must analyze the information and categorize it in order to identify the correct response. In many professions, analysis of signs (facts) may lead to a diagnosis (inference). Multiple-choice items that incorporate interpretive exercises (scenarios, drawings, tables, graphs, etc.) are useful in assessing analysis-level questions. A multiple-choice question that requires analysis of the distinction between facts and inferences is:

Which of the following is an inference explaining why North Carolina selected the cardinal as the state bird?

A. The cardinal is indigenous to the state.
B. The state's cardinal population is high.
C. The cardinal lives in all parts of the state.
D. The cardinal would look pretty on the state flag.

Competent professionals must be capable **of distinguishing between relevant and irrelevant information** when solving problems. Correct solutions require the candidate to attend to data that pertain to the problem and ignore information that is irrelevant or inconsequential. A multiple-choice question that requires analysis of the relevance of certain information using an interpretive exercise is:

Use the following information about plumbing parts and prices to answer Questions X through Y:

Component	**Quantity**	**Unit**	**Material**	**Labor**
8.2 gallon water cooler including piping, wall mounted	1	each	$352	$108
Type DWC copper tubing	5	linear ft.	$8	$40
Wrought copper tee, 2" diameter	1	each	$5	$34
Copper P trap, 1" pipe	1	each	$7	$14
Galvanized steel pipe, 2" diameter	4	linear ft.	$17	$27
Type L copper tubing, 3/8" diameter	10	linear ft.	$6	$46
Standard coupling for C.I. soil type, 4" diameter	5	each	$15	$99
Wrought copper, 90 degree elbow	3	each	$2	$33
Wrought copper Tee for solder joints	1	each	$1	$17
Copper stop and waster, 3/8" diameter	1	each	$3	$10

What is the total cost estimate for installation of an electric, self-contained, wall-hung water cooler with the relevant components, including overhead and profit at 14%?

A. $686
B. $782
C. $844
D. $962

Providing competent service requires the professional **to examine a situation and understand the relationship among its components.** An example of a multiple-choice question that requires analysis of the structure of the material presented and the relationships among its components is:

Which of the response options is the indirect object of the following sentence?

As president, Clinton gave Congress a clear mandate to address reform of the health care system.

A. Mandate
B. Congress
C. Reform
D. System

The Validation Process

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

When reviewing and validating examination questions, you will be asked to consider questions such as:

- Is mastery of the knowledge tested by the item essential for competent practice?
- 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 certification-level practitioner?
- Is the keyed answer correct?
- Can the keyed answer be defended if necessary?
- Are the distractors incorrect but still plausible?

You will also be asked to rate each item on scales such as importance, criticality, and frequency. Sample scales and instructions for use are provided below:

Rate the importance, criticality, and frequency of the tasks within each domain below, using the same scales that were used in rating the domains.

Importance	Criticality	Frequency
0 – Of No Importance	0 – No Harm	0 – Never
1 – Of Little Importance	1 – Minimal Harm	1 – Rarely
2 – Moderately Important	2 – Moderate Harm	2 – Sometimes
3 – Very Important	3 – Substantial Harm	3 – Often
4 – Extremely Important	4 – Extreme Harm	4 – Repeatedly

INSTRUCTIONS: Use the following ratings when judging each question in terms of importance, criticality, and frequency for the regulatory affairs professional.

Importance: How important is the knowledge tested by this question to the competent performance of the certified professional?

0 = Of No Importance
1 = Of Little Importance
2 = Moderately Important
3 = Very Important
4 = Extremely Important

Example: *It is important that we adhere to best practice in our professions.*
(Doing so helps to ensure that we do well for our clients and other stakeholders.)

Criticality: To what extent can the lack of knowledge tested by this question cause harm to the public and/or the profession? (Harm may be seen as physical, psychological, emotional, legal, financial, etc.)

0 = No Harm
1 = Minimal Harm
2 = Moderate Harm
3 = Substantial Harm
4 = Extreme Harm

Example: *It is critical that workers on high-rise buildings maintain a grip on the hammers. (Failure injures the public walking below, and impacts other stakeholders such as employers, insurers, etc.)*

Frequency: How many times does the certified professional perform duties requiring the knowledge noted in this question?

0 = Never
1 = Rarely
2 = Sometimes
3 = Often
4 = Repeatedly

Helpful Hints for Item Writers

Draw on your professional experience.

Think about professional problems or dilemmas you encounter on a regular basis. Alternately, think of situations you encounter less frequently, but whose outcomes have a critical effect on the functioning of your organization. Focus on problems where solutions require a combination of factual knowledge, professional judgment, and decision-making abilities. Pinpoint the knowledge and skill sets critical to successful job performance in these situations.

Locate your references in advance.

Each question you write must be substantiated by a published reference. This reference must be considered a standard, widely recognized text within your professional discipline. It would also be a good idea to review these texts in advance of your item writing session and to mark the pages that you think would provide useful background information for questions. Have these references alongside you as you write.

Write the stem.

Be *specific* and *concise*. More words do not equal more quality. When writing recall questions that test a candidate's knowledge of facts and principles, try not to address more than one fact or one principle per question item. When writing scenario questions, include only those details necessary for answering the question. Use active voice and avoid negatively worded stems (such as "Which of the following is NOT") whenever possible, as these can be tricky for test-takers.

Write the correct response.

The correct response should be one that a highly competent professional could formulate mentally after reading or hearing the question. It must be clearly worded and reflect widely accepted professional practice. It should be unique in comparison to the other answer options; no question should have more than one correct response listed.

Write the distractors.

Again, drawing on your professional experience and knowledge, write the incorrect responses to the questions. These "distractors" should be wrong, but plausible. You might begin by thinking about common professional myths, urban legends, and misunderstandings. Also, think about solutions that might work in situations similar to (but different from) the one stated in your question. The distractors should be answer options that a highly competent professional would rule out immediately, but that a minimally competent or incompetent professional would have to think about. Make sure the distractors and the correct answer to the question are grammatically consistent with each other and relatively equal in length.

Classify the question and record your reference.

Determine where the question falls on your group's classification system. Be as specific as possible. Record this classification on your item writing document along with your reference information and your name.

ITEM CONSTRUCTION FORM

TODAY'S DATE: _____

CASTLE Worldwide, Inc.
900 Perimeter Park Drive, Suite G
Morrisville, NC 27560

Phone: 919.572.6880
Fax: 919.361.2426
Email: info@castleworldwide.com

ITEM STEM:

CORRECT ANSWER

A.

B.

C.

D.

REFERENCE

TITLE _____

AUTHOR _____

PUBLISHER _____ DATE _____ PAGE _____

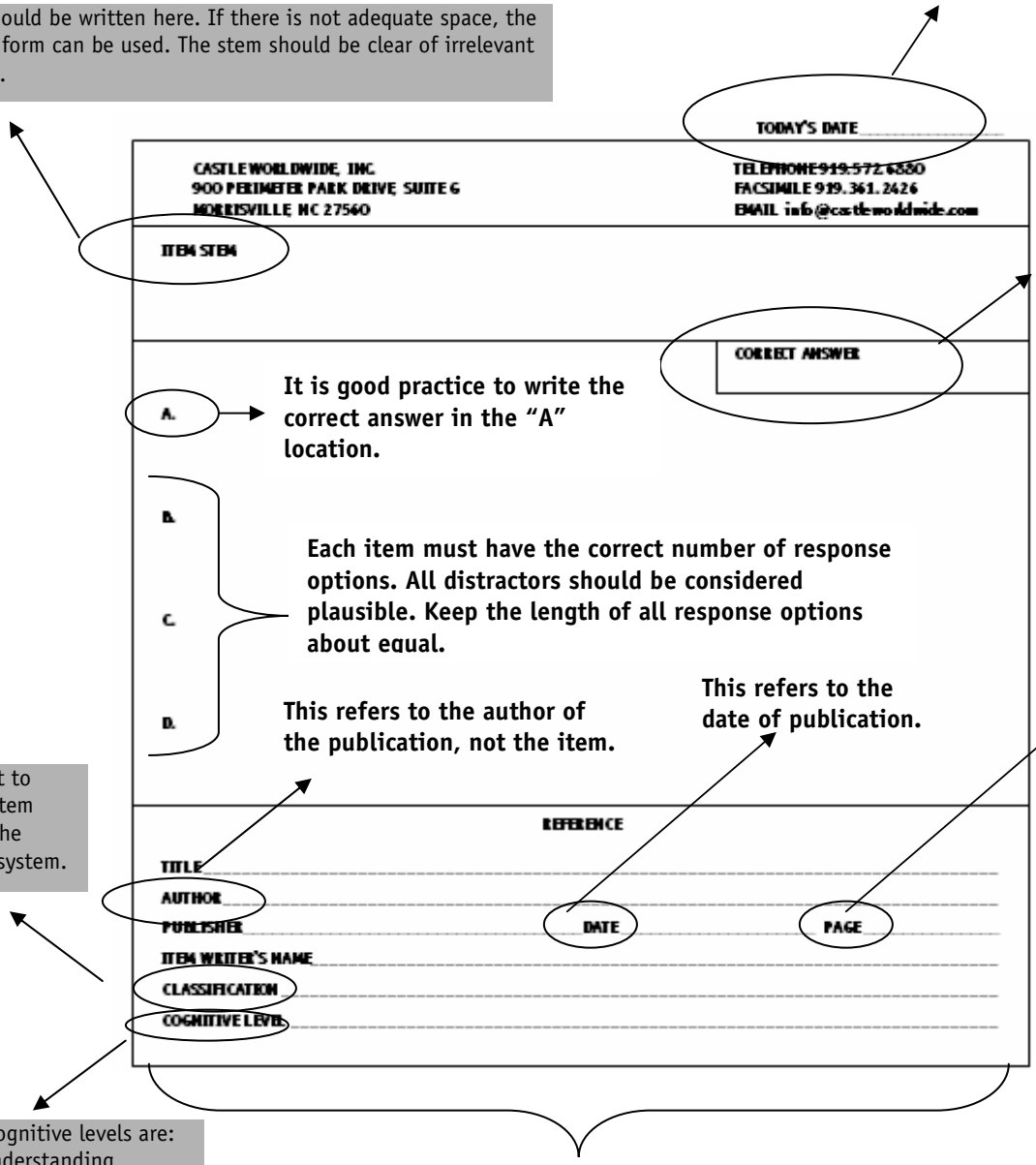
ITEM WRITER'S NAME _____

CLASSIFICATION NUMBER _____

COGNITIVE LEVEL _____

This refers to the date the item was constructed.

The stem should be written here. If there is not adequate space, the back of the form can be used. The stem should be clear of irrelevant information.



The Key, a vital piece of information, belongs here.

It is good practice to write the correct answer in the "A" location.

Each item must have the correct number of response options. All distractors should be considered plausible. Keep the length of all response options about equal.

This refers to the author of the publication, not the item.

This refers to the date of publication.

Don't forget the page number.

It is important to classify each item according to the classification system.

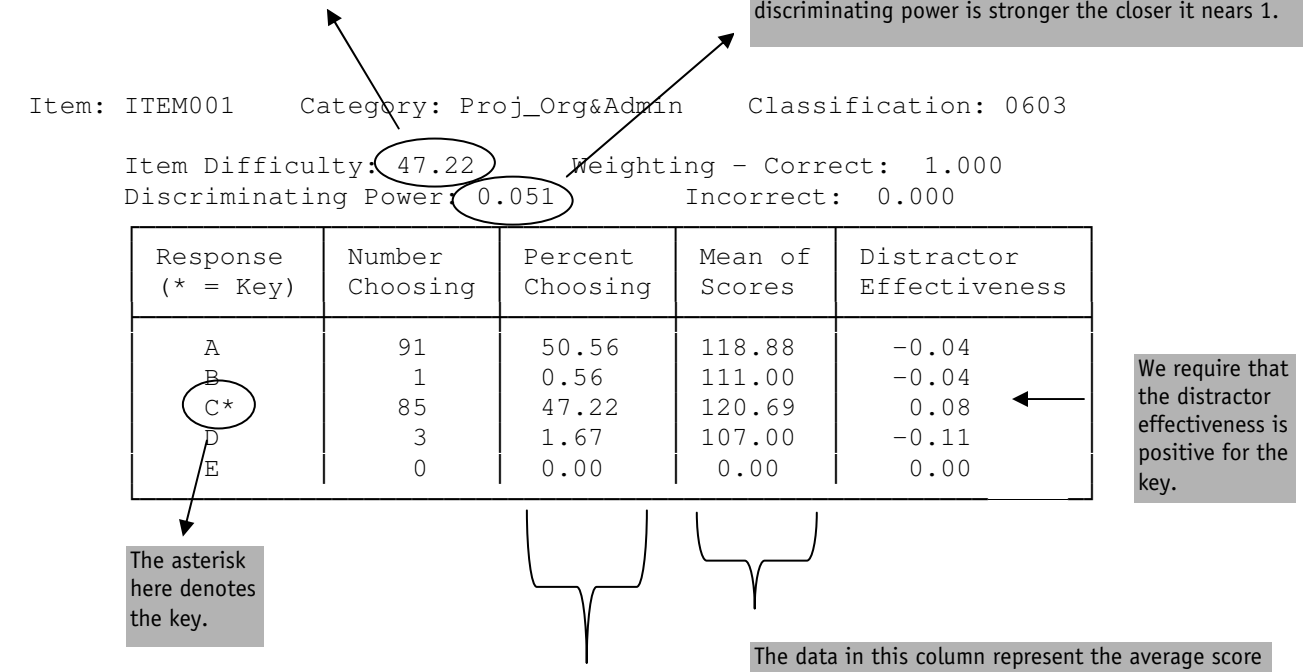
The three cognitive levels are:
I: Recall/Understanding
II: Application
III: Analysis

When writing new items, make sure you include a published reference where the content of each question may be verified. No items can be submitted without them.

Item Analysis Explanation

The item difficulty is the percentage of candidates who answered the question correctly. The recommended range for item difficulty set forth by CASTLE Worldwide, Inc., is between 92.00 (easy) and 30.00 (difficult). However, there are some instances where ratings slightly above 92.00 or below 30.00 are acceptable.

The discriminating power is a rating that denotes whether or not the question can discriminate between those candidates who possess the minimally acceptable level of knowledge to become certified and those candidates who do not. CASTLE's recommendation for the discriminating power is a positive rating at or above 0.15. The discriminating power is stronger the closer it nears 1.



Item: ITEM001 Category: Proj_Org&Admin Classification: 0603

Item Difficulty: 47.22 Weighting – Correct: 1.000
Discriminating Power: 0.051 Incorrect: 0.000

Response (* = Key)	Number Choosing	Percent Choosing	Mean of Scores	Distractor Effectiveness
A	91	50.56	118.88	-0.04
B	1	0.56	111.00	-0.04
C*	85	47.22	120.69	0.08
D	3	1.67	107.00	-0.11
E	0	0.00	0.00	0.00

We require that the distractor effectiveness is positive for the key.

The asterisk here denotes the key.

When the data in this column are lower than 3%, the accompanying distractor might not be plausible. These particular data are suspicious because there were more respondents that chose distractor A than the key. A key verification and distractor B and D revision is necessary in this situation.

The data in this column represent the average score that individuals who chose this response option received on the overall examination.

Item: ITEM008 Category: Document Classification: 0501

Item Difficulty: 69.49 Weighting - Correct: 1.000
Discriminating Power: 0.305 Incorrect: 0.000

Response (* = Key)	Number Choosing	Percent Choosing	Mean of Scores	Distractor Effectiveness
A	25	14.12	111.68	-0.22
B*	123	69.49	122.38	0.28
C	16	9.04	113.50	-0.13
D	13	7.34	115.85	-0.07
E	0	0.00	0.00	0.00

The statistics for this particular item are considered acceptable for the following reasons:

- The difficulty is neither too high or too low
- The discriminating power is substantially higher than 0.15
- No response options were chosen less than 3% of the time
- A large majority of examinees chose the correct answer
- The distractor effectiveness for the key is positive

NOTE: There is no distractor “E” with this item.

Item Sensitivity Review

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?
- Does the item use controversial or inflammatory content?
- Does the item use demeaning or offensive content?
- Does the item use stereotypical occupations (e.g., Chinese launderer) or stereotypical situations (e.g., boys as creative and successful, girls needing help with problems)?

Language

- Does the item use words with different or unfamiliar meanings?
- 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?
- Does the item contain inadequacies or ambiguities in the stem, keyed response, or distractors?

Glossary

Anchor Exam An examination form that sets the standard of passing for a given series of examinations.

Certification Authorized declaration validating that one has fulfilled the requirements of a given profession and may practice in the profession.

Classification System A systematic arrangement of examination content in groups or categories according to specified criteria. CASTLE Worldwide, Inc. uses a six digit coding system to represent the domain, task, and knowledge or skill a specific question covers.

Computer Based Testing (CBT) Refers to delivering examinations via computers. The examination questions are presented to candidates on a computer screen. Candidates choose their answers using the computer's mouse or keyboard, and their responses are recorded by the computer, rather than on an answer sheet.

Content Domain A body of knowledge, skills, and abilities defined so that items of knowledge or particular tasks can be clearly identified as included or excluded from the domain.

Cut Score A specified point on a score scale at or above which candidates pass or are accepted and below which candidates fail or are rejected. This is also sometimes called the passing score or passing point.

Discrimination The ability of a test or a test question to differentiate among qualified and unqualified individuals by measuring the extent to which the individual display the attribute that is being measured by the test or test question.

Distractor The options that are not correct answers. Distractors must be plausible; hence, they distract the less qualified test-taker from the correct answer.

Equating A process used to convert the score from one form of a test to the score of another form so that the scores are equivalent or parallel.

Equator Questions that are on all forms of an examination, including the anchor form. These questions are used to equate test forms.

Inter-rater Reliability Consistency of judgments made about candidates by raters or sets of raters.

Internet-Based Testing Computer-based testing. However, rather than send the examination to each testing center on computer media (compact disc), the examination is delivered via a secure, password-protected Web site. The examination and the candidate's answers are uploaded to the test provider's secure server. Test security is assured through configuration management, controlled loading, and availability.

Item A test question that consists of a stem, correct response, and distractors.

Item Analysis The process of assessing certain characteristics of test questions, specifically the question difficulty, the discrimination power, the candidates' mean scores, and the distractor effectiveness.

Item Difficulty The percentage of candidates answering a question correctly. This value can be computed to provide data about first-time candidates, retake candidates, ability level, etc.

Job Analysis Study Also known as role delineation study. The method of identifying the tasks performed for a specific job, or the knowledge, skills, and abilities required to perform a specific job.

Key The correct answer in a list of options.

Knowledge Statement An organized body of factual or procedural information is called knowledge**.**

Minimally Qualified Candidate An individual's competence in a particular job role can be seen as a continuum ranging (theoretically) from the complete lack of ability to the highest level of mastery. The term *minimum competence* suggests that the individual is capable of filling the role with sufficient mastery to not harm the public or the profession.

Options The list of possible answers for a question including the correct answer.

Performance Domain The major responsibilities or duties of a specific field of study. Each domain may be characterized as a major heading in an outline format and may include a brief behavioral description.

Psychometrics The design, administration, and interpretation of quantitative tests that measure psychological variables such as aptitude, intelligence, skill, and learning.

Raw score The unadjusted score on a test, usually determined by counting the number of correct answers.

Reliability The reliability of a test refers to the consistency of the test result. We interpret the reliability of a test as a measure of the likelihood that if we gave the test again under the same conditions, we would then observe the same scores.

Role Delineation Study Also known as job analysis study. The method of identifying the tasks performed for a specific job or the knowledge, skills, and abilities required to perform a specific job.

Scaled Score A score to which raw scores are converted by numerical transformation (e.g. standardized scores).

Score Any specific number resulting from the assessment of an individual. A number that expresses accomplishment either absolutely in points earned or by comparison to a standard.

Scoring Formula The formula by which the raw score on a test is obtained. The simplest scoring formula is the raw score equals the number of questions answered correctly.

Skill Statement The proficient physical, verbal, or mental manipulation of data, people, or objects is called skill. Skill embodies observable, quantifiable, and measurable performance parameters and may be psychomotor or cognitive in nature.

Standard Error of Measurement The standard deviation of the hypothesized distribution of test score means if multiple samples from which to compute the mean were available. We interpret the standard error of mean as a measure of variability we would observe in multiple sample or test administrations.

Stem The body of the question including any scenarios or qualifying information.

Subject Matter Expert A person with expertise in a given field or profession. Subject matter experts are used to develop the content of examinations.

Task Statement A comprehensive statement of work activity that elaborates upon the performance or content domain. Each task statement details a particular work activity in such a way that the series of task statements will offer a comprehensive and detailed description of each performance domain. In particular, task statements should answer the following questions:

- **What** activity do you perform?
- **To Whom** or **To What** is your activity directed?
- **Why** do you perform this activity?
- **How** do you accomplish the activity?

Test Specification A content outline that specifies what proportion of the test questions will deal with each content area.

Validation The process of rating each test question in order to determine how important, critical, and frequently the content tested by a specific question is used for a specific job.

Validity Refers to the quality of the inferences made from a test score/result. If the purpose of a particular examination is to certify a minimally qualified candidate in a particular profession, then the question we ask is whether minimal qualification can be inferred from the examination. Alternatively, validity can be conceptualized as the accuracy of the test score.

Traditional definition: Are we measuring what we intend to measure?

Weighted Scoring Scoring in which the number of points awarded for a correct response is not the same for all questions on a test.

Background Information on CASTLE Worldwide, Inc.

Founded in 1987 as Columbia Assessment Services, Inc., CASTLE Worldwide has grown into the sixth-largest certification and licensure testing company in the United States based on our ability to develop quality products and offer superior customer service to organizations of all sizes. CASTLE Worldwide is built on a history of developing quality high-stakes examinations for organizations that license or certify the expertise of professionals in fields ranging from medical technology to engineering to physical fitness. Tests can be developed for certification or licensure of professionals, to assess the knowledge and skills of employees for training purposes or as a pre-employment check for potential hires. CASTLE's psychometricians are trained to develop and validate assessment tools that measure whether or not a professional can perform his or her job-related duties with competence.

As a full-service company, CASTLE offers a broad array of products and services designed to meet the varied needs of business, educational, governmental, and professional organizations. Our hallmarks include high-quality, cost-effective development procedures; technological innovation and leadership; and our responsive customer service to clients and candidates. More than 50 organizations look to CASTLE Worldwide for assistance with their testing, training, learning, and certification programs.

CASTLE Worldwide offers its clients an extensive international network of test administration sites. Our secure Internet-based delivery system allows us to provide proprietary content directly to clients and their candidates any time of day, anywhere in the world. Internet-based practice tests and training programs make learning and studying convenient and self-paced. We offer our clients the expertise gained through years of service in leadership positions with the National Organization for Competency Assurance (NOCA), the National Commission for Certifying Agencies (NCCA), and the Council on Licensure, Enforcement, and Regulation (CLEAR).

CASTLE Worldwide's psychometric staff lends its expertise to leading industry publications including *Certification: A NOCA Handbook* and the revised *Standards for Educational and Psychological Testing*, published by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. CASTLE Worldwide is located in Research Triangle Park, North Carolina.

CASTLE worldwide
MULTIPLE CHOICES/ONE RIGHT ANSWER

CASTLE Worldwide, Inc. • 900 Perimeter Park Drive, Suite G • Morrisville, NC 27560