Item Writing and Review Manual

Nathan A. Thompson, Ph.D.
Jennifer P. Davis, Ph.D.

ASSESSMENT SYSTEMS
FOR GOOD MEASURE™

www.assess.com

Contents

1. Introduction

Purpose

Item writing is often regarded as an art form, but there is definitely a science to the process. The goal of this manual is to provide some simple tools to those responsible for developing new exam items or reviewing existing items. The key is to remember the end goal: to create an item that focuses on a piece of knowledge (or skill, ability, trait) and is able to differentiate between examinees with high and low levels of knowledge.

Terminology

The use of a standard terminology amongst subject matter experts (SMEs) participating in test development facilitates communication during the item writing and review process. The following provides a list of common terms and definitions.

Item – This is colloquially referred to as a test *question*. In many cases it is not a question, and therefore the more general term *item* is appropriate. For example, items can be statements with an agree/disagree response, sentence completion tasks, or essays.

Stem – This is the initial part of the item that is to be responded to. It refers to everything other than the options (answer), including things like reading passages, reference charts, and other prompts.

Prompt/Stimulus – In certain types of assessments, stems often contain a material other than simple question or statement text. For example, an item stem could contain a reading passage or audio of someone speaking, after which a specific item is presented.

Options – For multiple choice or multiple response items, a list of options is presented. These are also called alternatives, choices, answers, and distractors.

Key – The key is the option that you have marked as the correct answer. In rare cases where there are multiple correct answers, the key can identify the MOST CORRECT answer and the one for which examinees should receive credit.

Distractors – Every multiple choice item needs incorrect options. These are intended to distinguish candidates that know the material and candidates that do not. These are distractors as their purpose is to distract lower performing examinees from inadvertently selecting the correct answer.

Response – For a multiple choice item, this is the option that is selected. It might be an essay for a writing test, and spoken words for a speaking test. A “multiple response item” is one where an examinee can select more than one response, such as choosing the best two out of five options.

Rubric – This clearly defined set of criteria for scoring an item is set to a scale of numbers. It attempts to relate student responses on open response items to the standards and content of the test, providing a framework with which to evaluate responses.

2. Key Aspects of Item Development

A number of psychological and psychometric tenets underlie well-formatted exams and items:

Understanding of material

Item writers must have substantial knowledge of the constructs and material being examined. In other words, they must be of high ability themselves. Otherwise, how are they to write items that are able to identify examinees of high ability?

Rationale

Recording the reasoning behind the item and the correct response enables future reviewers and exam developers to understand the purpose for the item in its current state. The rationale can include a reference source such as a textbook page, or explanation of steps required to determine a solution.

Relevant content

An item writer is making the step from test specifications (outline or blueprints) to individual items. When writing an item, the learning objective or outline point for which the item is intended must be relevant to the overall goal of the test. The goal of test development is to cover each construct included in the exam outline, therefore newly written items must specifically map to the outline of the test.

Concise content

The item writer must focus on ensuring that the content of the item relates only to the piece of knowledge being assessed with no superfluous information included. An item is similar to a scientific experiment where all variables are held constant except for the variable being studied. This allows organizations to more accurately interpret the results of a test.

Understanding examinees

As items are developed and revised, it is important to understand the intended audience of the test and consider the perspective of potential examinees. If an item is to differentiate between low and medium examinees, the item writer must conceptualize what constitutes a low and medium examinee. How would examinees of low, medium, and high ability read, interpret, and answer the item?

Scoring and rubrics

Scoring multiple choice items is relatively easy because the correct answer gets 1 point and the remainders get 0 points. The correct answer should be fully correct while the others are fully incorrect or unassailably less correct (if that is the goal of the item). When developing open response questions (i.e. speaking, essays, short answer) or innovative items (i.e. simulations,

mock codes, performance exams), it is important to envision the rubric, or scoring system. If the question involves a conversational speaking response for an English test, how would you algorithmically assign points on a scale of 0 to 5? Multiple rubrics could be used for a single item, for instance spelling, sentence structure, and comprehension may all solicit a single point in an essay response.

Validity

The most important aspect of test scores is the validity of their interpretations. Items need to specifically measure what is supposed to be measured (called a *construct* in psychometrics), because then the scores on the test will accordingly reflect the construct (*construct-relevant variance*) and not any unrelated traits or aspects of the testing process (*construct-irrelevant variance*). This claim is supported by a chain of evidence from score interpretations back to the construct of interest. In the case of professional competency examinations, that chain is: job analysis – specifications – items – scores – interpretations. The test development process should revolve around ensuring linkage within the chain and documenting the linkage as much as possible.

Reliability

Exam scores should not vary from one delivery to the next. For instance, a group of a candidates taking a test today should not see significant differences in their score were they to take a similar examination next month. The quality of items, how they are developed and the content they contain, can influence a test's reliability.

3. General Tips

Formatting

There are two common formats for multiple choice items. The first is to simply ask a question and list several possible answers, the second is to formulate the item as a sentence completion or fill-in-the-blank task. Below is an example of the same item, with the same key and distractors, formatted in two different ways:

Format 1:

1. What is the capital of Norway?
    A. Oslo.*
    B. Bergen.
    C. Stavanger.
    D. Stockholm.

Format 2:

1. The capital of Norway is
    A. Oslo.*
    B. Bergen.
    C. Stavanger.
    D. Stockholm.

Avoid clues and cues

Avoid giving clues and cues for the test taker in the item text. These can be quite subtle. For instance, the correct option might begin with a vowel while the distractors begin with consonants.

Original:

1. An _____ is a large land mammal.
    A. elephant *
    B. whale
    C. shrew
    D. kangaroo

Revised:

1. A(n) _____ is a large land mammal.
    A. elephant *
    B. whale
    C. shrew
    D. kangaroo

Be sure to proofread exams for any instances that have an article (a/an) proceeding a word blank. Excluding (n) while having an answer beginning with a vowel could cause a test-wise examinee to eliminate an otherwise great distractor.

Candidates can also extrapolate clues across items. Verify that the stem text of one item does not give away the correct answer to another item.

Avoid negatives, qualifiers, and absolutes

Negatives (such as *not* or *except*) cause unnecessary confusion and can make items inappropriately difficult, penalizing examinees that are fast readers. The addition of a qualifiers (such as *best* or *worst*) implies that the item is referencing opinion rather than fact, therefore reducing the strength of the item. While sometimes accurate, absolutes should be avoided because there might be unique or rare unanticipated cases that invalidate the question.

Avoid All of the Above *and* A and B only

Response options including *all of the above* (*AOTA*) or some combination of aforementioned answers (e.g., *A and B only*) can cause unnecessary confusion. If used sparingly, the option itself can be a clue, for a test-wise examinee knows that the item writer is likely including *AOTA* because it is the correct answer.

Below are a few approaches one might use to rearrange an *AOTA* item. The same rules hold true for *none of the above*.

Original:

1. Which of the following is a city in Florida?
    A. Miami
    B. Tampa
    C. Jacksonville
    D. All of the Above *

Revised, make all other answers incorrect:

1. Which of the following is a city in Florida?
    A. Milwaukee
    B. Tampa *
    C. Phoenix
    D. Chicago

Revised, make options combinations to include more information:

1. Which of the following are cities in Florida?
    A. Miami and Tampa *
    B. Tampa and Milwaukee
    C. Chicago and Milwaukee
    D. Miami and Chicago

Revised, flip the stem to include options and vice versa:

1. Miami, Tampa and Jacksonville are all cities in _____.
    A. Alabama
    B. Florida *
    C. Illinois
    D. Wisconsin

Check grammar and punctuation

Item stem text should flow logically into the options listed below. For instance:

Original:

1. The capital of Kentucky is:
    A. Lexington
    B. Louisville
    C. Frankfort*
    D. Bowling Green

Revised, capitalize proper nouns and make complete sentence:

1. The capital of Kentucky is _____.
    A. Lexington
    B. Louisville
    C. Frankfort*
    D. Bowling Green

Be concise

Keep stems as short as possible. Be clear and concise and provide no extra information—especially information that could impact an examinee's response.

Original:

1. The capital of Kentucky was selected as a political compromise for being nearly equidistant from the two flagship cities of the state. The name of the capital is _____.
    A. Lexington
    B. Louisville
    C. Frankfort*
    D. Bowling Green

Revised, use only relevant information:

1. _____ is the capital of Kentucky.
    A. Lexington
    B. Louisville
    C. Frankfort*
    D. Bowling Green

Also avoid idioms and heavily specific jargon when possible. The use of acronyms is generally frowned upon. If a particular acronym is commonly used in the field (CPR for instance) use the full name once, with the abbreviation referenced in parentheses, before using the acronym on its own. For example: *cardiopulmonary resuscitation (CPR).*

Emphasize principles and concepts

Good items should not ask candidates to recall trivial facts. Avoid asking for specific numbers unless absolutely necessary and applicable to the constructs being tested.

Bad Example:

1. In the 2010 census, Madison, WI, had a population of _____.
    A. 233,209
    B. 243,394 *
    C. 227,221
    D. 239,874

If this test was being delivered in a setting in which recalling the number of residents in Madison in 2010 was relevant, the item writer might alter the stem to include "about" and provide estimates that vary enough to be distinguishable to highly skilled examinees. Otherwise, an item of this type should be replaced with one that evaluates more critical information.

Evenly distribute correct answers

Some item writers make the mistake of recording the correct answer first, thus making A the correct answer for every test. If you do this, be sure to use the *FastTest* setting that allows you to randomize the order in which options appear to examinees.

Keep distractors relevant

When creating distractors, remember that they can affect difficulty as much as the stem and content. Distractors that are misspelled or off topic can make the item too easy; whereas distractors that are too similar can make it too hard—especially if the similarity is not construct relevant.

For example, the following item tests the examinee's ability to spell rather than identify Vermont's capital.

1. The capital of Vermont is _____.
    A. Montpelier *
    B. Mountpelier
    C. Montpeleir
    D. Mountpeleir

Keep distractors consistent

Having options that vary in an irrelevant way causes unnecessary confusion. Try to keep distractors approximately the same length and structure as the key.

Original:

1. Define *hangar*.
    A. fame
    B. rough
    C. a storage area, like a garage, for a plane *
    D. calm

Revised to have similar length distractors:

1. Define *hangar*.
    A. complex; difficult to solve
    B. narrow-minded, a prejudiced person
    C. a storage area, like a garage, for a plane *
    D. taking credit for someone else's ideas

Avoid repetition

Avoid repeating certain words in each option by putting them in the stem text instead. This reduces the reading required for each item, makes the differentiating options more apparent, and slows examinee fatigue.

Original:

1. The front of a ship
    A. is called the bow.
    B. is called the stern.
    C. is called the port.
    D. is called the beam.

Revised:

1. The front of a ship is called the _____.
    A. bow
    B. stern
    C. port
    D. beam

Consider your demographic

Think about the content of your items. Would any subgroup be potentially disadvantaged in trying to answer this question? For instance, avoid asking questions about topics that are more common in certain areas, such as regional animals.

Logically decide and order numerical distractors

Ideally, numerical distractors should be distinct (not overlapping) and ordered in a logical manner.

Original:

1. The appropriate range of air pressure for an average road bike is _____ pounds per square inch.
    A. 20-40
    B. 90-120 *
    C. 30-50
    D. 50-70

Revised, no longer overlapping and ordered numerically:

1. The appropriate range of air pressure for an average road bike is _____ pounds per square inch.
    A. 20-50
    B. 50-90
    C. 90-120 *
    D. 120-150

Avoid personal pronouns

Referring to the test taker, by using the pronoun “you”, implies that the candidate is making a choice rather than selecting an answer based on fact.

Original:

1. Half way into a long hike you notice that you have a blister on the top of your toe. You take off your boots and socks and
    A. pop the blister, drain the fluid, and continue hiking.
    B. ignore the blister until you get home.
    C. apply a loose bandage or moleskin to protect the area. *
    D. Peel away the blistered skin and cover the area with a bandage.

Revised to solicit a medical best practice rather than personal preference:

1. Which of the following do medical professionals recommend that hikers do when they encounter a non-weight bearing blister mid-hike?
    A. Pop the blister, drain the fluid, and continue hiking.
    B. Ignore the blister until the hike is completed.
    C. Use a loose bandage or moleskin to protect the area. *
    D. Peel away the blistered skin and cover the area with a bandage

Be consistent

Some item writers choose to emphasize key aspects of item stem text. If you choose to place emphasis on certain words, be sure that the emphasis is consistent throughout all items. Don’t underline some words and bold others.

4. Item Review

Item writers often feel relief after finishing an item writing assignment. However, that is just the first step in a long process of ensuring the quality of each item and the validity of resulting scores. Items should be reviewed by at least one other SME in order to ensure validity from a quality control perspective. Once items have been reviewed sufficiently to minimize possible challenges from examinees, items can be compiled into a test and pilot tested with examinees.

After a test form has been seen by a sufficient number of examinees, it is statistically analyzed to flag items with possible issues. These items are then reviewed again to ensure legitimate content. Items that are modified are saved as a new version because they are a “new” item from a psychometric perspective.

Below is a checklist, recommend for use in the initial review of an item by a SME other than the item writer.

Review Checklist
Item content maps to specific learning objectives and test relevant constructs
Each item assesses a key principle rather than a trivial fact (even if the fact is technically relevant)
Items are posed in the positive, avoiding NOT and EXCEPT
The key is correct and there is only one correct answer
Distractors are feasible, but definitely incorrect
No clues or cues are given in the stem text (a/an) or across the test
Grammar and punctuation are correctly utilized, sentences flow properly
Idioms, unnecessary jargon, and acronyms are not present
Item stem is concise and includes no irrelevant information
Options do not include repetitive text
Options are of similar length and format
Essay or short answer rubrics adequately evaluates the candidate's ability
Content is equally relevant across all groups and does not disadvantage any groups
Items fall within the difficulty scope of the intended test
Record the rationale (why the key is the correct answer) and a source if applicable
Options are logically ordered and numerical guidelines are followed
Correct answer is evenly distributed across the options; the key is not always A
Items are formatted to solicit facts rather than opinions