Item Writing and Review Guide

Nathan Thompson, PhD
CEO & CPO

March, 2021

ASC

Table of Contents

Want to learn more about ASC's psychometric services?

solutions@assess.com

http://assess.com

651.383.4311

Introduction

Item writing is often regarded as an art form, but there is a science to the process. The key is to remember the end goal: that the item will focus on a piece of knowledge (or skill, ability, trait) and differentiate between examinees with high and low levels of knowledge. There are several important components of a good item writer.

Validity

The most important aspect of test scores is the **validity** of their interpretations. We need the item 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.

Goals in Item Writing

The following list discusses important goals to keep in mind while authoring items.

Mapping objectives

An item writer is making the step from test specifications (outline or blueprints) to individual items. Because the goal of test development is to link the steps as closely as possible, newly written items must specifically map to the outline of the test. Therefore, when writing an item, the specification (learning objective, outline point) for which the item is intended must be recorded.

Clear and concise

The item writer must be focused in making sure the content of the item only relates to the piece of knowledge, with no superfluous information, and the examinee responses are designed to differentiate among examinees. An item is similar to a scientific experiment, where all variables are held constant except for the variable being studied, so that it might be evaluated more accurately.

Understanding examinees

Therefore, one of the most important aspects is to think like an examinee. How would examinees of low, medium, and high ability read, interpret, and answer the item? Obviously, we want

examinees of higher ability to be able to recognize the correct response. Examinees of lower ability will not be able to, but not necessarily because they are confused. This in turn means that the item writer must fully understand the intended audience of the test. Items should be of appropriate difficulty; if an item is to differentiate between low and medium examinees, the item writer must conceptualize what constitutes a low and medium examinee.

Anticipating scoring

Also, writers should keep in mind the scoring of the item. For multiple choice items, this is 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. Unfortunately, this simplification does not lend itself well to assessing complex constructs such as higher-order thinking or psychomotor skills. Therefore, when developing open response questions (speaking, essays, short answer) or innovative items (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?

Understanding material

Perhaps the most obvious, and therefore possibly overlooked, requirement is that item writers have substantial knowledge of the material. 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?

Record rationale

Record the reasoning behind the item and the correct response. If the item is used and reviewed a year from now, whoever is reviewing will want to know the rationale. The rationale can include a reference source such as a textbook page, or explanation of steps required to determine a solution.

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*, but 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, essays, and real-life performance tasks.
Stem – This is the initial part of the item that presents what is to be answered. 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 assessment, stems often contain a prompt or stimulus. For example, the stem could contain a reading passage or audio of someone speaking, after which a specific question is presented.

Options – For multiple choice or multiple response items, a list of options is presented. These are also called alternatives, choices, and answers.
Key – The correct option.
Distractors – The incorrect options.
Response – The response of the examinee to an item. 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 – A clearly defined set of criteria for scoring an item 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.

Multiple choice items

Multiple choice items are the most common type of test item. There is a reason for this: not only are they simple to write, but they are easily scored in an objective manner, making them highly reliable. By comparison, "innovative" items such as like speaking and simulations have more fidelity to the real world, but are difficult to score objectively.

There are two common formats for multiple choice items. The simpler format is to simply ask a question and list several possible answers.

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

The second format is to formulate the question as a sentence completion or fill-in-the-blank.

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

Normal grammatical rules should be followed. Since the answers all represent the last word in a sentence, they are followed by periods (full-stops). In the first format, the stem has a question mark because it is a question, while the options are all standalone phrases and therefore have no punctuation. Similarly, the stem in the sentence completion does not have a colon, as there would be

no colon at that point in a normal sentence, and the options would not have capital letters (except that they do here because they are proper nouns).

In an effort to reduce construct-irrelevant variance, items should be formatted as similarly as possible. All items should have the same font style and size, same number of options when feasible, and similar writing style/structure.

Polytomous Items

Polytomous items are those that have multiple point levels. These are colloquially referred to as *partial credit* but it is best practice not to have partial points; instead, the scoring should be consecutive integers so that the items can be scored with item response theory models such as the generalized partial credit model.

Much of the points covered above will also pertain to polytomous items. The anticipation of scoring and understanding of examinees become more important and complex; instead of the item being scored correct/incorrect, you might now have point values of 0,1,2,3,4 and need to determine answers that represent distinct ability levels of students – which is much harder than it looks.

General tips

1. Avoid clues and cues for the test taker. These can sometimes be quite subtle. For instance, the correct option might begin with a vowel while the distractors begin with consonants. If the sentence completion ends with the word "an" this would tip off the examinees.

   Bad example:
   An _____ is a large land mammal.
   A. elephant
   B. whale
   C. shrew
   D. kangaroo

2. Avoid "NOT" items, "All of the above," and "A and B only" items, as they can be confusing.

   Bad example:
   Which of the following cities is NOT in Texas?
   A. Dallas
   B. Houston
   C. Little Rock
   D. San Antonio

3. Check grammar and punctuation.

   Bad example:
   The capital of Kentucky is:
   A. Lexington

B. Louisville
C. Frankfort
D. Bowling Green

4. Keep stems as short as possible; be clear and concise, with no extra information, especially information that could impact responses.

Bad example:
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

5. Emphasize principles rather than trivial facts.

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

6. Evenly distribute the key among possible locations.

Bad example: *tending to make A the correct answer because the item writer will first think of the correct answer and then create distractors.*

7. Remember that distractors can affect difficulty as much as the key. They should be relevant. Irrelevant distractors make the item too easy; distractors too similar to the key make the item too difficult, especially if the similarity is construct-irrelevant.

Bad example:
The capital of Vermont is _____.
A. Montpelier
B. Mountpelier
C. Montpeleir
D. Mountpeleir

8. Options should be of similar length/format.

Bad example:
The longest river system in North America is the _____.
A. Columbia
B. Mississippi-Missouri
C. Colorado
D. Yukon

9. Avoid “always” and “never” because there might be rare unanticipated cases.

Bad example:
In New York, it never snows in June.

TRUE
FALSE

10. Avoid options that naturally group together.
    Bad example:
    Travel faster than the speed of sound is called _____.
    A. supersonic
    B. subsonic
    C. sonic
    D. stratospheric

11. Avoid repetition of text in each option – put it in the stem if possible, so it is used only once.
    Bad example:
    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.

12. Avoid idioms and other limited-use or esoteric terminology.
    Bad example:
    The two companies negotiated all day regarding small points of the contract, because they each considered the final price to be the
    A. thousand pound gorilla in the room
    B. best part of the day.
    C. easiest thing to negotiate.
    D. icing on the cake.

13. Think about the content: would any subgroup be potentially disadvantaged?
    Bad example: *on national exam, asking questions about topics that are more common in certain areas, such as medical exam asking about scorpion bites.*

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 before the item is 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.

Baranowski (2006) mentions how many studies find that items with slight flaws are not much more difficult than without the slight flaw. For example, an implausible distractor does not make the item less correct (the key is still fine), just slightly easier. The reason for the review process is to ensure

validity from a quality control perspective. Irrelevant information can lead to possible challenges by examinees that the item is not valid, namely not focusing on measuring what is supposed to be assessed.

Appendix A presents the checklist used in the initial review of an item by a SME other than the item writer.

Appendix A: Item Review Checklist

When reviewing an item written by someone other than yourself, please use the following checklist. If reviewing a number of items, please paste this table into Microsoft Excel and copy the second column as many times as needed.

Verification	**Check (✓) if OK, or describe recommendation**
Item content correctly maps to specifications or learning objectives	
Key is correct	
Distractors are feasible but definitely incorrect	
No cues for examinee (such as a/an)	
Does not include “NOT” , “All of the above,” “A and B only,” “always,” or “never”	
Check grammar and punctuation; if sentence completion, should flow appropriately.	
Stem has no superfluous information.	
Item assesses key principle, not trivial fact (even if the fact technically falls under blueprints)	
No idioms or esoteric jargon	
Options are of similar length and format	
No options naturally group together.	
No repetition of text to begin each option	
If rubric-scored, verify that the rubric adequately evaluates with regard to intended interpretation	
Content is equally relevant across major groups, not disadvantaging any group	