Item Writers Manual

Preparing Examination Items

Written by
Leon J. Gross, Ph.D.
Director of Psychometrics and Research
National Board of Examiners in Optometry®

Foreword

This online manual, prepared by Dr. Leon Gross, Director of Psychometrics and Research of the National Board of Examiners in Optometry®, is an updated version of a manual that was originally published in 1980. At that time, it was one of the initiatives of the National Board intended to assist item writers develop and prepare excellent examination questions. In addition to assisting in improving the quality of item writing for our own examinations, the manual was also designed to help readers in their own educational and/or evaluation endeavors. Over the years it has been in continuous use by faculty members and practitioners as the basis of the item data bank which provides the building blocks of the "National Boards". It has been translated into French, German and Spanish, and is being used by faculty members in other countries to help them adopt multiple choice questions (MCQs) as the method of developing standardized tests for future optometrists.

The National Board is committed to providing the very best national evaluation mechanism for the profession of optometry, and since this manual was first published the Board has moved on many fronts to address the issue of competence measurement in clinical practice. This includes the development of the Part III (Patient Care) examination, which includes other methods of assessing clinical competence, in addition to MCQs. The current "National Boards" are an integral part of competence assessment in optometry, and your involvement in the process is essential and is very much appreciated.

Please utilize the information and techniques outlined here by Dr. Gross, so that together we can shoulder the responsibility of providing the finest national evaluation of new doctors of optometry entering practice - to serve the public.

Thank you for your efforts and your contributions.
Arol R. Augsburger, O.D., M.S., President
National Board of Examiners in Optometry®

1 What is an objective item?

Objective test questions or items are those in which the examinee must select the one correct or best response from two or more alternatives. Objective items may be administered in a variety of formats (e.g., multiple-choice, matching, true-false); however, psychometric research suggests that multiple-choice (MC) items are the most reliable and valid of the objective formats. Each of the National Board tests of cognitive skill is composed entirely of MC items; this manual is designed to facilitate proper and effective MC item development.

2 What should test items measure

Each item should focus on a single clear concept or problem that is related to a central and essential issue at the appropriate point in the examinee's career. Thus, an item written for Part I should reflect the skills and abilities of a second-year *student*, rather than those of a colleague or practitioner. It is therefore not valid to develop items that test for esoterica or minutiae, for knowledge that would be "nice" for an optometrist to have but that is not particularly important, or for knowledge of recent faculty research accomplishments. Since National Board examinations are administered nationally, each item that you submit should similarly be targeted for the national optometry student population and *not* give the students at any one school or locale an unfair advantage in responding. Finally, items should *not* be written to trick the examinee or to insure that a particular text or article has been read. The issue of what items should measure is amplified in Section 7 dealing with item taxonomy.

3 Basic item writing considerations and terminology

Objective test questions are properly referred to as *items* because they may be stated in the form of a question or an incomplete statement which is completed by the selected answer. The incomplete statement or question to which the examinee must respond is known as the *stem*, and the choices (as in multiple-choice) from which the examinee must respond are known as *alternatives*, *responses*, or *options*; those that are incorrect are referred to as *distractors*. The typical item will contain 4 or 5 options, depending primarily on distractor plausibility. Thus, as an item writer, it is entirely up to you to decide for each test item whether it should contain 4 or 5 options; however, you should *not* exceed the *maximum* of 5 alternatives.

4 Formulating the item stem

As mentioned in the preceding section, the stem can be in the form of a complete question or an incomplete statement. Neither form is uniformly preferable to the other; preference is dictated by ease, simplicity, and clarity of wording. In the example below, the sentence completion format is superior. The stem is the most important portion of an item, for it presents the task or problem to the examinee; it therefore must be clearly written in order that examinees know what is expected of them.

Example 1

Question Format	Sentence Completion Format
Which of the following terms may be referred to as an objective test question?	Objective test questions are known as:
a. item b. alternative c. response d. distractor	a. items b. alternatives c. responses d. distractors

If the item requires the examinee to read the options before the nature of the item can be comprehended, then the stem is inadequate. To illustrate the dilemma this poses for the examinee, consider the following example.

Example 2

Poor	Improved
Vitamin D:	Vitamin D is known to prevent
a. prevents scurvy b. is an antibiotic c. is found in milk d. is also known as niacin	a. scurvy b. pellagra c. rickets d. night blindness

You should note in the poor item that: (1) the abbreviated stem does not convey the nature of the problem to the examinee; and (2) there is a lack of homogeneity in the options. Items with abbreviated stems (such as this) are poor from the examinee's vantage point and frequently indicate that the item writer has not properly formulated for him/herself what (s)he wishes to test (e.g., use, prevalence, or classification of vitamin D). The improved item is also testing knowledge of vitamin D; however, this item has a well-focused stem and homogeneous responses. It is clearly superior to the former item and is more likely to contribute to test validity and reliability. Consistent with providing the examinee with a clear understanding of what is expected, the stem should not be structured for fill-in responses. Such "breaks" in the stem unnecessarily impede clarity. An example of this is illustrated next.

Example 3

Poor	Improved
The profession of _____ is concerned primarily with vision care.	The profession primarily concerned with vision care is:
a. optometry b. podiatry c. cardiology d. pharmacy	a. optometry b. podiatry c. cardiology d. pharmacy

One final note on the importance of the stem is that it largely determines the types of cognitive skills that the item measures. For example, one item may test for *recall* of knowledge while another tests for *application* of knowledge. This important issue will be discussed at length in Section 7.

5 Generating distractor options

Although the development of the item stem is critical for the item to contribute to a valid measure of achievement, the construction of distractors is equally critical but probably more difficult. The most obvious purpose of a good distractor is, in a sense, to camouflage the correct response. As such, a good distractor will be attractive or compelling to the examinee who does not know the correct response or who has incomplete or superficial knowledge of the material. While distractors need not be equally compelling, they should represent an array of errors or misconceptions for a range of seriousness or crudeness. Thus, while each distractor will be an incorrect response, some will be more "acceptable" than others in terms of their degree of seriousness/crudeness. One very useful way of generating distractors is to anticipate how examinees could independently arrive at logical though incorrect responses, and then to utilize such incorrect responses as distractors.

Example 4

Poor	Improved
15 x 5 =	15 x 5 =
a. 73 b. 74 c. 75 d. 76	a. 3 b. 20 c. 55 d. 75

As an example of this approach, consider the example above. The distractors in the item on the left are designed solely to camouflage the correct response. However, since there is no obviously rational way to compute the responses in options (a), (b), and (d), the distractors are ineffective despite their numeric proximity to the keyed response. In contrast, the corresponding item on the right contains distractors that bear a rational though incorrect logic. Option (a) would be obtained by solving the problem through division rather than multiplication; option (b) involves the erroneous use of addition; and option (c) represents the use of multiplication but without properly carrying the 2 from the "ones" to the "tens" column. This item illustrates a very important principle in generating distractors: treat distractors as a codification of incorrect responses that examinees would likely supply if multiple choices were not provided.

As such, do not feel compelled to aim for a uniform number of options per item; the number of plausible incorrect alternatives will vary from item to item. However, you should not exceed five options (i.e. one correct, and four distractor options).

There are two additional considerations for writing distractors. First, as was mentioned in an earlier section, you should not try to "trick" the examinee. Therefore, do not write distractors with ambiguous or misleading terms. Second, your items should not contain combination options such as "none of the above," "all of the above," and "a and b above." These options contain inherent psychometric weaknesses that lower test validity and the reliability of the results.

6 Editorial considerations

It is important that you review and edit items before submission. This will lower the incidence of: (1) ambiguities or other problems preventing the high-achieving student from responding correctly; and (2) cues that "give away" the correct response to a low-achieving student. Editorial issues that you may find useful are outlined below.

A. Options should be independent and mutually exclusive. Option (d) in the following example overlaps with options (a) and (b) and creates response difficulties.

Example 5

Poor	Improved
The charge of the National Board of Examiners in Optometry is:	The charge of the National Board of Examiners in Optometry is:
a. certification testing b. accreditation c. lobbying d. credentialing	a. certification testing b. accreditation c. lobbying d. research dissemination

B. Items asking for a particular course of action should ***not*** solicit the student's opinion (What would you do?) but rather, what the ***best*** action is (What ***should*** you do?). Some clinical encounter items might require a specific action ***sequence***. Consider the next example.

Example 6

Poor	Improved
A patient comes to your office complaining of blurred vision and headaches while wearing corrective lenses. What would you do?	A patient comes to your office complaining of blurred vision and headaches while wearing corrective lenses. What should you do FIRST?
a. administer a vision test b. obtain patient history c. examine patient's lens prescription d. refer patient to a neurologist	a. same options b. c. d.

C. Avoid using double negatives, since they merely serve to confuse the examinee. The problem this poses is illustrated below.

Example 7

Poor	Improved
Not being able to exert full control over one's voluntary muscles is not characteristic of:	The ability to exert full control over one's voluntary muscles is characteristic of:
a. paralysis b. trembling c. rigidity d. coordination	a. paralysis b. trembling c. rigidity d. coordination

D. Avoid using absolute qualifiers (e.g., always, never, totally, completely) in the options, since they do not allow for the exceptional case and therefore, usually provide a cue that the response is incorrect.

Example 8

Poor	Improved
Which of the following BEST describes the effects of excessive doses of chlorpromazine?	Which of the following BEST describes the effects of excessive doses of chlorpromazine?
a. The patient's coordination will always be impaired. b. The patient will be totally incapacitated. c. The patient's handwriting will sometimes be affected. d. The patient will have complete loss of tactile sensation	a. The patient's coordination will usually be impaired. b. The patient will be almost totally incapacitated. c. The patient's handwriting will sometimes be affected. d. The patient will have almost complete loss of tactile sensation

E. Overly long, qualified and technical options are typically cues to the correct response. Insure that each option within a given item is similar in form and length.

Example 9

Poor	Improved
When taking antibiotic drugs, the patient should:	When taking antibiotic drugs, the patient should:
a. not exercise	a. not exercise until three days after medication ceases
b. expect a temporary period of remission	b. expect a 2- to 6- week period of remission
c. remain in a state of quarantine	c. remain in a state of quarantine for five days
d. continue the medication even after symptoms and discomfort are no longer present, because if there is a residual infection, the infectious organism may develop a tolerance for the drug and excessive doses may subsequently be necessary	d. continue the medication after symptoms and discomfort vanish

F. Options that are not grammatically consistent with the stem are typically incorrect. Therefore, insure that each option provides grammatical consistency.

Example 10

Poor	Improved
Fungi that are commonly eaten are:	Fungi that are commonly eaten are:
a. apple	a. apples
b. tomato	b. tomato's
c. spaghetti	c. spaghetti
d. mushrooms	d. mushrooms

7 Taxonomy levels

In addition to testing for achievement in specific content areas, multiple-choice items measure general cognitive processes. These thought processes or taxonomy levels represent a hierarchical method for categorizing the cognitive skills needed to correctly answer an item. While this hierarchical paradigm was originally based on six distinct levels, experience in health professions testing suggests that the taxonomic model is easier to use when compressed into the following three levels:

- 1 - recall of knowledge
- 2 - simple calculation and data interpretation
- 3 - problem solving and evaluation

A description of these levels appears below.

Level 1: Recall of Knowledge refers to the remembering of previously learned material either by recognition or rote recall. This may involve the recall of a wide range of material, from specific facts to complete theories; however, recall is all that is required to demonstrate mastery.

Level 2: Simple Calculation and Data Interpretation refers to the ability to compute values (e.g., a mean score), interpret data (e.g., whether an obtained test result exceeds a critical value), or recognize the general significance of verbal or pictorial data (e.g., recognizing instrumentation data that are out of range).

Level 3: Problem Solving and Evaluation refers to the ability to utilize verbal or pictorial data for the resolution of a problem or evaluation of a situation (e.g., diagnosing the visual problem of a patient on the basis of clinical data).

Since the latter two levels measure the student's ability to *apply* previously learned knowledge in a clinical context, test items written at these levels are particularly critical in order for certification/licensure examinations to be considered *job relevant* or *competency based,* rather than merely a composite regurgitation of isolated facts. While the ability to recall knowledge is often important to test, it is *in*sufficient for demonstrating overall professional competency, for it does not tell us as examiners that the student can *apply* that knowledge. In measuring promotional or professional readiness, the demonstrated application of knowledge is the essence of test validity.

Writing items at taxonomy levels 2 and 3 is challenging and time consuming, but it is also rewarding and not as difficult as it initially appears to be. Rather than testing for facts, principles, or characteristics of the visual system, provide the student with data such as charts, tables, diagrams, or kodachromes, or perhaps write an item stem that places the student in a clinical situation in which you provide patient symptoms and/or examination findings which must be evaluated.

The following examples taken from other health professions fields illustrate how an item measuring a particular concept at the tax 1 level may be rewritten to measure the same concept at the tax 3 level. You are strongly encouraged to emphasize tax levels 2 and 3 in your item writing; however, it is recognized that the clinical optometric content areas will lend themselves to this more readily than the theory-oriented areas. Nonetheless, all content areas do contain some concepts that can be tested in a tax 2 or 3 mode; you should try to accomplish this whenever possible.

Example 11

Concept tested: Recognition of Psychological Defense Mechanisms

Taxonomy 1	Comment
Projection is a defense mechanism that is MOST characterized by: a. expressing the opposite of one's true sentiment b. attributing one's true sentiment to another person c. a conscious denial on one's true sentiment d. an avoidance of individuals expressing sentiments parallel with one's true sentiment	In this form of the item, the examinee need only recall characteristics of the defense mechanism. a. paralysis b. trembling c. rigidity d. coordination

Taxonomy 3	Comment
A particular individual who had just entered college, had quickly begun to experience a longing for his parents and strong feeling of insecurity. After one week at school, he began to inform several of his dormitory mates that THEY were probably insecure about being away from home. This student's behavior exemplifies the defense mechanism known as: a. reaction formation b. repression c. denial d. intellectualization e. projection	In this form of the item, the examinee must recall characteristics of the defense mechanism AND apply these characteristics in evaluating the behavior presented.

Example 12

Concept tested: Recognition of Toxoplasmosis

Taxonomy 1	Comment
Blindness, brain damage, and death will MOST likely result from: a. hepatitis b. coronary thrombosis c. toxoplasmosis d. pleurisy	In this form of the item, the examinee need only recall the outcomes of the disease.

Taxonomy 3	Comment
Several researchers at a particular hospital were investigating the case histories of three patients. During early adulthood, Patient A had become blind, Patient B had developed a neurological dysfunction, and Patient C had suddenly died. Their case histories revealed the following common characteristics: (1) their mothers had pet cats; (2) their fathers were alcoholics; and (3) each patient was quite obese. On the basis of these data, the most logical conclusion that can be drawn is that all three individuals would **MOST** likely have been victims of:	In this form of the item, the examinee must recall both the causes and outcomes of the disease and evaluate the relevant from the irrelevant information in order to identify the disease in question.

8 Final checklist

1. Does the item concern an important aspect of the subject? Avoid esoterica, minutiae, provincial interests, etc.

2. Does the item stem present a single, definite, important task?

3. Is the item presented more effectively as a question or as an incomplete statement?

4. Is the language of the item as direct, relevant, and devoid of excess wording as possible?

5. Is the item keyed correctly? Is there a *single* correct or best answer?

6. Will the student understand the focus of the item without having to read the responses?

7. Are the options similar in form, length, and grammatical construction with the stem and with each other?

8. Are the options plausible and attractive to the low achiever; that is, do the distractors attract students who have incomplete or superficial knowledge of the material? Insure that you have *not* used an ambiguous or misleading word or phrase just to "trick" the student.

9. Are the responses independent and mutually exclusive?

10. Does the item use less-than-absolute qualifiers (e.g., usually, sometimes, rarely) instead of absolute qualifiers?

11. Does the stem ask the student what course of action (s)he "should" take rather than the course of action (s)he "would" take? Should an *initial* course of action be requested?

12. Are your items oriented toward measuring data interpretation (tax 2) and problem solving skills (tax 3)?