Guidelines for Item Writing

The New England College of Optometry

Overview

This guide summarizes a set of evidence-informed principles for writing high-quality assessment items. The principles outlined here are drawn from the *Standards for Educational and Psychological Testing* and established assessment literature. These same standards underpin the development of high-stakes assessments such as those produced by the National Board of Examiners in Optometry (NBEO), the Optometry Examining Board of Canada (OEBC), and represent widely accepted practices for producing valid, reliable, fair, and defensible examinations.

Four Characteristics of High-Quality Assessment

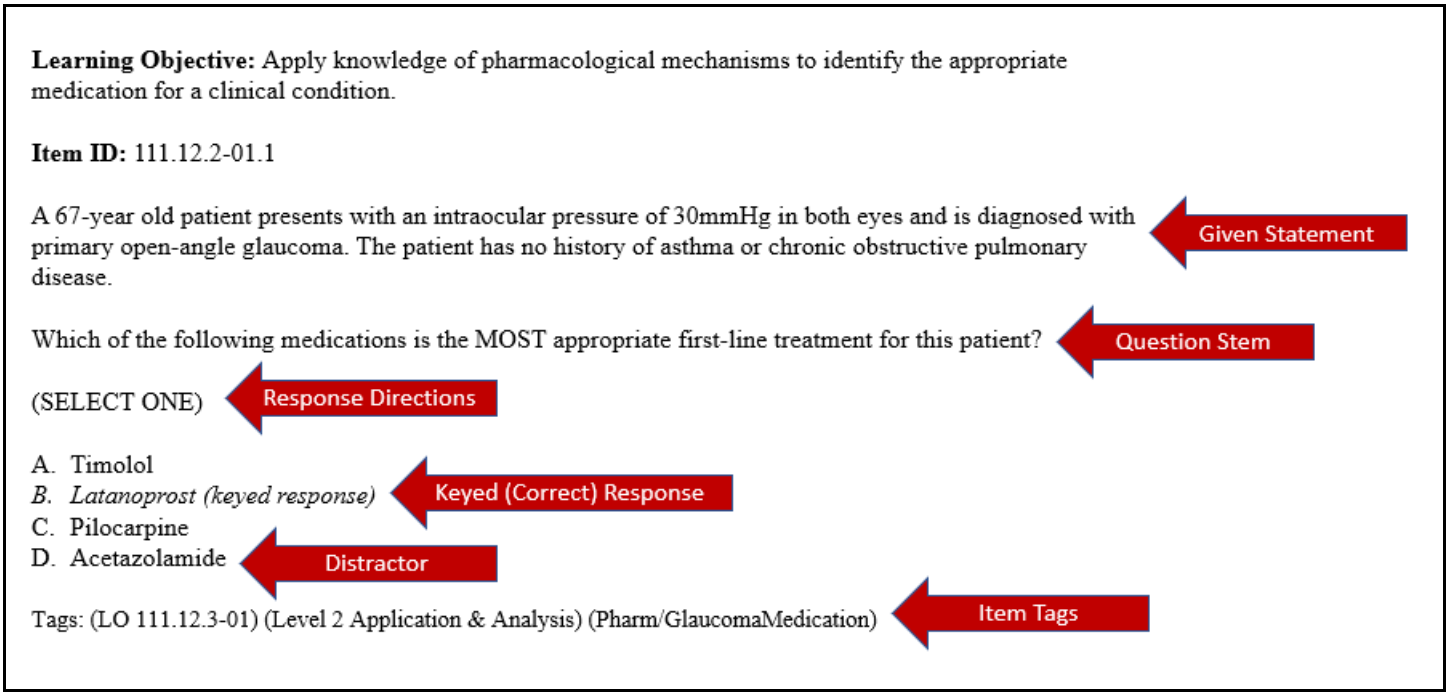
Validity refers to the degree to which an assessment measures the knowledge, skills, or clinical judgment it is intended to measure. Assessment items should align with the intended learning objective, appropriate cognitive level, and course content while avoiding construct-irrelevant factors that affect difficulty, clarity, or interpretation.

Reliability refers to the consistency of assessment results. Clear, unambiguous items with well-functioning response options help ensure that student performance reflects differences in knowledge rather than differences in how questions are interpreted.

Fairness means that every student has an equal opportunity to demonstrate what they know. Assessment items should use clear, inclusive language, avoid unnecessary complexity or bias, and measure only the intended construct.

Discrimination refers to an item's ability to distinguish between students who have mastered the learning objective and those who have not. High-quality items are answered correctly more often by knowledgeable students than by less-prepared students. Clear stems, plausible distractors, and appropriate cognitive demand all contribute to strong item discrimination.

Anatomy of an Item



Learning Objective	Defines the specific knowledge, skill, or clinical judgment the item is intended to measure. The learning objective serves as the foundation for the item and guides the selection of the appropriate item type and cognitive demand.
Given Statement	Provides the background or contextual information necessary to answer the question (e.g., a clinical vignette, laboratory findings, image description, or other relevant data). Include only information that contributes directly to the construct being assessed.
Question Stem	Presents the problem students are expected to solve. The stem should be written as a clear, concise, complete question (or directive for matching and some constructed-response items) that measures a single learning objective.
Response Directions	Instructs students how to respond to the item (e.g., Select One, Select Two, Select All That Apply). Directions should clearly communicate the expected response format before students evaluate the answer options.
Keyed Response	The correct answer or answers to the item. The keyed response should be unambiguously correct, directly aligned with the learning objective, and supported by current evidence and course content.
Distractors	Incorrect response options designed to be plausible to students who have not yet mastered the learning objective. Effective distractors reflect common misconceptions or errors while remaining clearly incorrect to well-prepared students.

Item Types

The choice of item type has a direct impact on what an assessment measures and how well it measures it. Different item formats place different cognitive demands on students, vary in their ability to discriminate between levels of performance, and differ in their reliability, efficiency, and scoring characteristics.

For most assessments, traditional multiple-choice items provide the strongest combination of validity, reliability, and discrimination, and should therefore serve as the primary item type. Alternative formats, such as multiple-response, matching, and short-answer items, have appropriate uses but should be reserved for situations in which they provide a better measure of the intended learning objective.

Approved Item Types

Item Type	Rationale
Multiple-choice (single correct response)	The preferred format for most summative assessments. Well-written multiple-choice items efficiently measure a wide range of knowledge and cognitive skills while providing high reliability, objective scoring, and strong psychometric performance. They generally demonstrate the strongest discrimination and reliability of any item type.
Multiple-response (multiple correct responses)	Because these items increase cognitive demand and difficulty, they should be used only when recognizing multiple correct responses is essential to the construct being assessed. *NOTE: 10-15% of NBEO items are multiple-response; OEBC does not employ multiple-response items.*
Matching	Efficient for assessing relationships between related concepts, terms, images, diagnoses, or treatments. Matching items are particularly useful when multiple items share a common set of response options. Matching items generally perform more poorly than multiple-choice items but similarly to multiple-response items. *NOTE: Neither NBEO nor OEBC employ matching items*
Short Answer (constructed response)	Appropriate when students must generate rather than recognize a response, such as performing calculations, interpreting findings, or providing a brief explanation. They generally discriminate well and are reliable when crafted carefully. *Note: Constructed-response items are used on NBEO part 3 and OEBC part 2 examinations.*

Discouraged Item Types

Item Type	Rationale
True/False	True/False items provide a 50% chance of answering correctly by guessing and they tend to measure recognition rather than meaningful understanding. Universally, they demonstrate poor discrimination and reliability. True/False items are generally disallowed by most professional testing organizations, including NBEO and OEBC.
Multiple True/False (K-type)	Items that require students to evaluate multiple statements (e.g., "A and C are correct") increase unnecessary cognitive complexity, encourage test-taking strategies rather than content mastery, and generally demonstrate poorer psychometric performance. These formats are generally disallowed by most professional testing organizations, including NBEO and OEBC.

Cognitive Demand

Every assessment item should measure (and be tagged with) the level of thinking described by the learning objective. The cognitive demand of an item should be intentional and should reflect what students are expected to know or be able to do at that point in the curriculum. At NECO, learning objectives and assessment items are classified using a three-level cognitive demand framework adapted from Bloom's Taxonomy:

Level	Description
Level 1: Knowledge & Understanding	Recall, recognize, define, describe, or explain facts, concepts, or information.
Level 2: Application & Analysis	Apply knowledge to interpret or analyze information, solve problems, or differentiate critical from unimportant information in a given context.
Level 3: Evaluation & Synthesis	Integrate information, prioritize alternatives, justify decisions, or determine the most appropriate course of action.

Recommended Distribution of Cognitive Complexity

The distribution of cognitive demand across an examination should reflect the primary purpose of the course and the intended learning outcomes. Foundational courses are expected to emphasize knowledge acquisition and comprehension, whereas clinically focused courses should place increasing emphasis on application and clinical reasoning. The following distributions are intended as general guidelines rather than fixed requirements:

Course Type	Level 1	Level 2	Level 3
Foundational (e.g., basic science, introductory)	60-80%	20-40%	0-10%
Systems-based (e.g., disease, conceptual)	30-50%	40-60%	10-20%
Clinical application	10-25%	40-50%	35-50%
Advanced Clinical Decision-making	0-10%	30-40%	50-70%

Learning Objectives

A well-defined learning objective serves as the foundation for high-quality assessment. It guides the selection of the appropriate item type, informs the intended level of cognitive demand, and provides evidence that the assessment is measuring what students were expected to learn. Conversely, items that are not grounded in clearly defined learning objectives tend to be less valid and reliable, and are prone to construct-irrelevant variance.

If your course does not have clearly defined learning objectives, or if your existing objectives do not fully reflect the intended outcomes of the course, use the following questions to guide the development of each assessment item:

- What specific knowledge or skill am I trying to measure?
- What level of thinking do I expect students to demonstrate? Is it sufficient to recall or explain a concept, or should students be able to apply, analyze, evaluate, or make a clinical decision using that knowledge?
- What evidence would convince me that a student possesses this knowledge or skill?
- Does the item as written measure the intended knowledge or skill directly, or is it introducing unnecessary difficulty unrelated to the learning outcome?

Tagging Items

Item tagging provides a systematic way to organize assessment items and analyze exams over time. Tags allow faculty to evaluate exam blueprints, monitor coverage of course content, identify gaps or redundancies, and assemble future forms more efficiently. They also allow for data aggregation to evaluate student performance on specific exam areas or topics.

At minimum, each assessment item should be tagged with:

- **Learning Objective:** The code for the learning objective that the item measures
- **Cognitive Demand:** The intended cognitive demand level of the item (Level 1, Level 2, Level 3)
- **NBEO Topic:** The relevant NBEO content topic or subtopic, when applicable

Faculty are encouraged to add additional tags that support their own organization or instructional needs.

Question Stem Construction

Complete Question

The question stem should be phrased in the form of a clear and direct question. The stem should present a single, well-defined problem and include only the information students need to answer the question correctly.

Matching items are an exception to this rule, as the stem should be phrased in the form of a directive (e.g., "Match each condition with the most appropriate diagnosis") rather than posing a question.

Short answer items may be an exception to this rule, as well, depending on what is being assessed. The prompt may either ask a question (e.g., What is the product of…") or provide a directive (e.g., "In the space provided, explain the mechanism of action of atropine".

Avoid Fill-in-the-Blank Format

Avoid using fill-in-the-blank question stems. This format interrupts the natural flow of reading, encourages sentence completion rather than problem solving, and may introduce grammatical or syntactic cues that are unrelated to the knowledge or skill being assessed. In many cases, fill-in-the-blank items can also make the question less direct by obscuring exactly what students are being asked to do.

Whenever possible, write the stem as a complete question that clearly defines the problem students are expected to solve. Complete-question stems are generally easier to read, reduce opportunities for cueing, and more closely reflect the item formats used on high-stakes professional examinations.

Avoid Cueing or Teaching in the Stem

The purpose of the question stem is to define the problem students are expected to solve, not to provide information that leads them to the correct answer. Avoid including unnecessary explanations, definitions, diagnostic clues, clinically irrelevant information or instructional language that teaches the concept being assessed or narrows the range of plausible responses. Every sentence in the stem should contribute directly to the construct being measured.

Likewise, avoid wording that unintentionally cues the correct response through repeated terminology, distinctive phrases, or excessive emphasis. If students can answer the question by recognizing clues in the stem rather than applying their knowledge or reasoning, the item is measuring test-taking skill rather than the intended learning outcome.

Clinical Scenarios or Background Information

When a clinical scenario or other contextual information is necessary, present it first as a brief statement prior to the question. Organizing the stem in this way helps students distinguish relevant background information from the task they are being asked to perform, reduces extraneous cognitive load, and promotes a more consistent item structure across examinations.

Avoid Addressing the Test Taker

Write assessment items in the third person and avoid directly addressing the student using words such as “you” or “your”. Second-person stems can subtly change the frame of reference for the test taker, introducing construct irrelevant variance. It shifts the test taker’s cognitive processing, creating a distraction. Using third-person language standardizes interpretation, reduces ambiguity, and maintains a consistent professional tone.

Demographic Information

Include demographic information such as age, sex, gender identity, race, ethnicity, occupation, or other personal characteristics only when it is directly relevant to the knowledge or clinical decision being assessed. Including irrelevant demographic details can introduce unnecessary reading burden, reinforce stereotypes or implicit bias, and inadvertently cue or mislead students. If removing a demographic characteristic does not change how the question is answered, it likely does not belong in the item.

Use Inclusive and Unbiased Language

Question stems should be written using clear, inclusive, and professionally appropriate language that allows all students an equal opportunity to demonstrate their knowledge and skill. Avoid language or contextual details that introduce construct-irrelevant variance, bias, or that unfairly disadvantage students.

Avoid stereotypes, generalizations, or assumptions that associate demographic characteristics with particular behaviors, conditions, or outcomes unless they are supported by the clinical context and essential to the construct being measured.

Likewise, avoid slang, colloquialism, idioms, regional expressions, culturally specific references, humor, or unnecessary jargon that may be unfamiliar or distracting. Use terminology that is widely accepted within the profession and ensure that any abbreviation or acronyms are commonly recognized by the intended audience.

Avoid Negative Phrasing

Questions that ask students to identify what is not true or to select an exception (e.g., NOT, EXCEPT, LEAST likely, INCORRECT) are negatively phrased items. These formats require students to mentally reverse the meaning of the question or identify the one option that does not belong, increasing reading demands and the likelihood of errors unrelated to the knowledge being assessed.

Item writers sometimes use negatively phrased questions because they assume these items demonstrate that students recognize all of the correct alternatives. In reality, they provide evidence only that the student identified the incorrect option. Students may arrive at the correct answer through partial knowledge or logical elimination without demonstrating mastery of every correct alternative. As a result, negatively phrased items often measure less of the intended construct while introducing unnecessary cognitive complexity.

The *Standards for Educational and Psychological Testing* (followed by the NBEO, OEBC, NBME, and other professional testing organizations) recommend using negatively phrased items **sparingly** and only when recognizing an exception, contraindication, or incorrect action is itself an intended learning outcome.

Appropriate Use of Negatively Phrased Question Stem

Learning Objective: *Identify unsafe medication orders.*

Question Stem:
A patient's chart contains the following medication orders. Which order should **NOT** be implemented because it places the patient at significant risk of harm?

In the example above, recognizing an unsafe medication order is itself the competency being assessed. The purpose of the item is not simply to identify the correct treatment, but to ensure that students can recognize and avoid a potentially harmful clinical decision. Because this skill is essential for protecting patient safety, the negatively phrased stem directly reflects the intended learning outcome and provides a valid measure of clinical competence.

Negatively phrased stems should not be used when a positively phrased question can assess the same knowledge equally well. If negative phrasing is unavoidable, make the negative word highly conspicuous (e.g., NOT or EXCEPT in ALL CAPS or bold) to reduce the likelihood of misinterpretation.

“Best Answer” Format

Use "best answer" questions only when the learning objective requires students to evaluate multiple reasonable alternatives and determine the most appropriate course of action. This format is well suited for assessing clinical judgment, prioritization, and decision-making, where more than one response may be acceptable but one is clearly preferable given the clinical context.

When only one response is objectively correct independent of the clinical context, ask students to identify that response directly rather than asking for the BEST, MOST APPROPRIATE, or MOST ACCURATE option. These qualifiers should be reserved for situations in which students are expected to compare multiple potential alternatives. If no meaningful comparison is required, the qualifier adds unnecessary complexity without improving the validity of the assessment. For example:

> **Not Appropriate:** Which of the following is the MOST ACCURATE description of the optic chiasm?
>
> **Preferred:** Which of the following correctly describes the optic chiasm?

Directions

Provide clear response directions immediately following the question stem in parentheses. Explicit directions help ensure students understand the expected response format before reading the answer options, reducing confusion and preventing avoidable response errors.

For traditional multiple-choice items with a single correct response, use (SELECT ONE).

For multiple-response items, specify the required number of correct responses using the following format: (SELECT 2), (SELECT 3), etc.

In most objective testing platforms, including the NBEO examinations, multiple-response items are scored as all-or-nothing. Students receive credit only if all correct responses are selected and no incorrect responses are chosen. Consequently, directions such as (SELECT ALL THAT APPLY) substantially increase item difficulty because students are not told how many responses are correct and receive no partial credit for partially correct answers.

For this reason, the NBEO does not use (SELECT ALL THAT APPLY) items. Unless there is a compelling assessment rationale, faculty are strongly encouraged to specify the number of correct responses in the directions.

For matching items, the directions should be detailed, indicate how many response options may be used, whether responses can be used more than once, and whether all responses will be used. Example: (Directions: Place the stages in the correct sequence, beginning with the first step. Each option will be used once.)

For short-answer items, the directions should tell students the expected type, scope, and format of the response. Clear directions improve scoring consistency and reduce ambiguity. Examples include:

- (Respond in one complete sentence)
- (Respond in 1-2 sentences)
- (Enter a numeric value only)

Capitalize Qualifiers

When qualifiers such as BEST, MOST, or LEAST are necessary, emphasize them using capitalization so they are not overlooked.

Terminology

Using standardized terminology improves clarity, promotes fairness, and ensures that items remain consistent with professional licensure examinations. Avoid the use of proper names, eponyms, abbreviations, acronyms, or institution-specific terminology unless they are essential to the construct being measured or reflect accepted professional usage. When referencing medications, pharmaceuticals, or other branded commercial products, use the generic names wherever possible.

Graphics, Media, and Images

If the item includes images, figures, videos, or other media, ensure they are directly relevant to the question and of sufficient quality to interpret accurately. All images, graphics, and media must be accessible to all learners (e.g., Alt text, screen reader compatible, etc.), free of any protected health information (PHI), and either copyright-free or used with appropriate permission.

Question Stem Checklist

- ☐ Written as a complete question ending with a question mark
- ☐ Clear, concise, and free of unnecessary information
- ☐ Uses third person language and avoids direct references to the test taker (e.g., “you”)
- ☐ Avoids cueing or teaching in the stem
- ☐ The question stem is positively worded if possible
- ☐ Measures a single construct or learning objective
- ☐ Any background info and/or clinical context (Given Statement) is presented before the question
- ☐ Demographic information included only when clinically relevant
- ☐ Question stem is free of distracting or biasing language (e.g., pronouns, racial epithets, slang)
- ☐ Qualifiers (e.g., BEST, MOST, LEAST, EXCEPT) used only when absolutely necessary and emphasized
- ☐ Generic terminology used whenever possible
- ☐ Images or media are relevant, high quality, accessible, and appropriately licensed
- ☐ Response instructions provide explicit directions (e.g., Select 2)

Constructing Response Options

Response options should be carefully constructed to support the measurement of the intended learning outcome without introducing unnecessary clues or ambiguity. All options should be plausible, grammatically consistent with the stem, and presented in a parallel format.

Number of Response Options

For traditional multiple-choice items, provide four to five response options, including the keyed response. Although research suggests that three-option items can be equally effective when all distractors are plausible, developing two highly effective distractors is often more challenging than it appears. Faculty should prioritize writing plausible, functioning distractors and avoid adding weak or implausible options simply to increase the number of responses.

For multiple-response items, include one more incorrect distractor than correct response options whenever possible, with a maximum of seven total response options (e.g., A “Select 2” question should have at least three distractors). This provides sufficient discrimination while minimizing unnecessary cognitive load.

Parallel Construction

Response options should be parallel in length, grammatical structure, tense, complexity, and level of specificity. Parallel construction prevents students from identifying the correct response based on formatting or wording rather than content knowledge.

Avoid repeating identical words or phrases across all response options. Whenever possible, place shared wording in the stem rather than repeating it in each option.

Avoid Cueing

Response options should not provide unintended clues to the correct answer. Avoid grammatical mismatches between the stem and the correct response, unusually long or detailed correct answers, or distinctive wording that makes one option stand out.

None/All of the Above

Avoid using response options such as All of the Above, None of the Above, or combination options (e.g., A and C, B and D only). These formats reduce the ability of an item to accurately measure student knowledge by allowing test-wise students to arrive at the correct answer through logic or process of elimination rather than content mastery.

For example, if a student can identify that two response options are correct, they may correctly select All of the Above without evaluating the remaining option. Conversely, selecting None of the Above demonstrates only that the listed options are incorrect, not that the student knows the correct answer. Combination response options require students to solve two problems simultaneously: determining the correctness of individual responses and then evaluating combinations of those responses. This increases cognitive complexity without improving the measurement of the intended learning outcome.

Whenever possible, write items with a single, clearly correct response or, when appropriate, use a multiple-response format that explicitly instructs students to identify all correct responses.

Universal Qualifiers

Avoid universal qualifiers such as *always*, *never*, *only*, and *all* unless they are unquestionably accurate. These words are often recognized by experienced test-takers as indicators of an incorrect response, making the item easier for reasons unrelated to the construct being assessed.

Keyed (Correct) Response(s)

The keyed response should be clearly and unambiguously correct based on current evidence and the content taught in the course. In single-response multiple-choice items, there should be only one defensibly correct answer. In multiple-response items, all keyed responses should be absolutely correct in the context of the question.

Avoid repeating distinctive words or phrases from the stem in the keyed response, as this may unintentionally signal the correct answer to students.

The correct response should not be noticeably longer, more specific, or more technically detailed than the distractors.

Distractors

Effective distractors are plausible responses that reflect common misconceptions, errors in reasoning, incomplete understanding, or incorrect application of knowledge. A distractor should be attractive to students who have not yet mastered the material while remaining clearly incorrect to well-prepared students.

Distractors should be mutually exclusive and should not overlap conceptually with one another or with the keyed response. Avoid humorous responses, vague statements, implausible alternatives, or options that can be eliminated immediately without applying course knowledge.

Like the keyed response, distractors should be grammatically consistent with the stem and parallel in length, structure, complexity, and style. Well-written distractors increase an item's ability to discriminate between students who have mastered the learning objective and those who have not.

Use Inclusive and Unbiased Language

Response options should be written using clear, inclusive, and professionally appropriate language that allows all students an equal opportunity to demonstrate their knowledge and skill. Avoid language or contextual details that introduce construct-irrelevant variance, bias, or that unfairly disadvantages students.

Avoid stereotypes, generalizations, or assumptions that associate demographic characteristics with particular behaviors, conditions, or outcomes unless they are supported by the clinical context and essential to the construct being measured.

Likewise, avoid slang, colloquialism, idioms, regional expressions, culturally specific references, humor, or unnecessary jargon that may be unfamiliar or distracting. Use terminology that is widely accepted within the profession and ensure that any abbreviation or acronyms are commonly recognized by the intended audience.

Response Option Checklist

- ☐ Response options are parallel in length, grammar, tense, and level of specificity
- ☐ Shared wording has been moved to the stem whenever possible
- ☐ Correct number of response options are present (MC: 3-5, MS: no more than 7 total)
- ☐ No “All of the Above” or “None of the Above” options
- ☐ No combination responses (e.g., A and C, B and D only)
- ☐ The keyed response(s) is clearly and unambiguously correct in the context of the question
- ☐ Every distractor is plausible but unequivocally incorrect in the context of the question
- ☐ Distractors reflect realistic misconceptions or common errors
- ☐ Distractors are mutually exclusive and do not overlap with one another
- ☐ Response options are free of unnecessary information
- ☐ Response options are free of distracting or biasing language (e.g., pronouns, racial epithets, slang)