Test/Quiz Item (Question) Design Guidelines

Based *mostly* from Popham, W. J. (2020). *Classroom assessment: What teachers need to know* (9th edition). Pearson.
Compiled by mjstuve@bsu.edu for my students in EDPS 345, 390, 393, 646, etc. • **Last update March 2026**

The General Process for evaluating an item's (question's) viability BEFORE using it in a test with students:

A. Assume every item is technically flawed; seek to prove that it is not.
B. Run it through its specific **item type guidelines (Section 3 and 4** on next page). Identify potential flaws.
C. Run it through the **Bias checks (Section 2** on this page). Identify potential threats to fairness (i.e. identify bias).
D. Run it through the **General Principles in Section 1.**
E. Prioritize the severity and kind of violation and address in order of severity. **Look for the flaws that could result in one of the three scenarios outlined in the FIOS box to the right.**
F. Revise accordingly.
G. Don't hesitate to start over with an item (see Principle 1.6)
H. Get another set of teacher eyes on the item.
I. **Repeat steps A-H to be more confident in your item.**
J. Then gather data with student assessment and conduct an Item analysis. Revise accordingly.

Flawed Item Outcome Scenarios (FIOS)

Your goal with judging item quality ***before*** conducting the test is to avoid one the three outcome scenarios of a flawed item:

1. The student who knows the answer to a question is prohibited from correctly answering it because of a flaw in the item (i.e. the **false negative** outcome)
2. The child who doesn't know the answer gets it correct because of a flaw in the item, not because of their ability or luck (i.e. the **false positive** outcome)
3. **Spill-over effects** (or collateral damage): Instrument flaws or errors incur anxiety, frustration, doubt, etc. which affects student concentration, motivation, mood, etc. to perform well.

1. **Section 1: General Principles** for Item Design
The first five are referred to as **"Commandments"** in Popham's Chapter 6, but here we call the **General Principles.** They apply all item types including constructed response items. The latter six principles are synthesized from Popham and other sources. The list now "goes to eleven".
   1.1. Avoid opaque (unclear) directions to students regarding the test's overall instructions.
   1.2. Avoid ambiguous statements in your items. Note, content **in**accuracy in the item can often look like this.
   1.3. Avoid unintentional clues within or across items that give away the correct response.
   1.4. Avoid complex syntax in items (i.e. sentence structure). Note, content **in**accuracy can often appear as this.
   1.5. Do not use vocabulary that is more advanced than the learner's reading level.
   1.6. Choose the most effective item type in relation to the question's depth of knowledge and related pedagogical factors of your classroom (McMillan, 2019; Russell & Airasian, 2012). Consider different item types.
   1.7. Validity and Item Difficulty. A question that is too superficial or relatively too easy does not afford meaningful inferences, even if the question has no other technical flaws.
   1.8. Content and Construct Validity. Make sure the question serves the learning objectives of the assessment. In other words, make sure that meaningful inferences about student learning are possible. Avoid questions that test something other than the content or constructs being sought/taught.
   1.9. Content Accuracy. Make sure your question is accurate to your content (not counting errors like 1.2 and 1.4 above).
   1.10. Reliability: Make sure the item is not vulnerable to generating inconsistent responses among students who know a correct or suitable response or across items that purport to measure the same concept (i.e. internal consistency).
   1.11. Fairness. Make sure each item addresses Absence of Bias in Section 2 on page 1.

2. **Section 2. Fairness (Absence of Bias)**
(From Popham Chapter 5; Woolfolk Chapter 15, McMillan Chapter 3). Make sure your test items DO NOT exhibit the following:
   2.1. Offensiveness: When an item contains stereotypes or other offensive, distressing, or anxiety-causing content.
   2.2. Unfair Penalization: When the item systematically disadvantages a student's ability to answer the question correctly because of factors associated with one or more group memberships (e.g. race, ethnicity, gender, SES, etc.).
   2.3. Disparate Impact: Be aware of ***potential*** factors that have been shown in standardized tests that correlate to identifiable groups of learners. Such findings ***might*** affect how you interpret group-based differential performance in your classroom assessments. This topic needs thorough and ongoing attention on your part as an educator.

Get updates to this document and other assessment resources at: https://espace.bsu.edu/lamp

See Item ***Type*** Guidelines (Sections 3 and 4) on Page 2 (on other side if printed double-sided)

3. **Section 3. Selected Response Items** (Popham Ch. 6; Woolfolk Ch. 15; McMillan Ch. 8/9)
   - 3.1. **Binary-Choice (B-C) Items:**
     - 3.1.1. Phrase items so that *superficial* (lack of thought) analysis by the student suggests a wrong answer.
     - 3.1.2. **Avoid** using negative statements (e.g. using "not" without visual emphasis). Never use double negatives.
     - 3.1.3. Include only one concept in each statement.
     - 3.1.4. Across the B-C items of a particular type, have an approximately equal number of items representing the two choices (i.e. the number of items where the correct answer is "true" is roughly the same as "false").
     - 3.1.5. Across the B-C items of a particular type (e.g. True/False), keep item length *roughly* similar for both options.
     - 3.1.6. Keep the statements succinct and significant; *avoid* lengthy or trivial statements.
   - 3.2. **Multiple Binary-Choice Items:**
     - 3.2.1. Separate item clusters visually from one another.
     - 3.2.2. Make certain that each item meshes well with the cluster's stimulus material.
   - 3.3. **Multiple-Choice (M-C) Items:**
     - 3.3.1. The stem should consist of a self-contained question or problem, i.e. the central idea to be tested.
     - 3.3.2. *Avoid* negatively stated stems but if you must use **NOT** (purposefully, say, to increase item difficulty), make it visually evident to the learner.
     - 3.3.3. Do not let the length or wording of the correct response(s) supply unintended clues. Common mistakes in making correct responses include (from McMillan, 2019):
       - 3.3.3.1. Significantly longer that incorrect ones
       - 3.3.3.2. Noticeably more elaborate or detailed or general or technical compared to incorrect responses
       - 3.3.3.3. Grammatically correct relative to the grammar of the stem
       - 3.3.3.4. Repeats the wording from the stem
     - 3.3.4. Randomly assign correct answers to alternative positions in the response options.
     - 3.3.5. Avoid "all-of-the-above" alternatives, but "none-of-the-above" can be used to increase item difficulty.
     - 3.3.6. Make sure the correct choice is exclusive, verifiable and discriminating (which, in assessment, means that there is a **detectable difference** between the correct choice and incorrect ones), unless…
     - 3.3.7. If the item is a multiple select variant of multiple choice (i.e. "check all that apply"), all the correct response options must be verifiable and discriminating compared to the set of incorrect responses options.
     - 3.3.8. All **distractors** should be plausible, not overly complex, and verifiably incorrect within the domain or topic being tested.
     - 3.3.9. Minimize the amount of reading in the response options.
   - 3.4. **Matching Items:**
     - 3.4.1. Employ homogeneous lists of responses and premises.
     - 3.4.2. Use relatively brief lists, placing the shorter words or phrases at the right.
     - 3.4.3. Employ more responses (e.g. terms to match to definitions) than premises (e.g. definitions).
     - 3.4.4. Order the responses logically.
     - 3.4.5. Describe the basis for matching and the number of times responses may be used.
     - 3.4.6. Formatting (visual presentation) and instructions matter (i.e. can affect the student's response):
       - 3.4.6.1. Be clear in the instructions and mechanics for the matching task
       - 3.4.6.2. Keep the stem and all premises and responses for an item on a single page.

4. **Section 4. Constructed Response Items (Popham Chapter 7; McMillan Chapter 9)**
   - 4.1. Item-Writing Guidelines for **Short-Answer Items**
     - 4.1.1. Usually employ direct questions rather than incomplete statements, particularly for young students.
     - 4.1.2. Structure the item so that a response should be concise.
     - 4.1.3. Place blanks in the margin for direct questions or near the end of incomplete statements.
     - 4.1.4. For incomplete statements, use only one or, at most, two blanks.
     - 4.1.5. Make sure blanks for all items are equal in length.
   - 4.2. Item-Writing and Scoring Guidelines for **Essay Items**
     - 4.2.1. Clearly convey the extensiveness desired (i.e. how much to write) and format for sufficient space.
     - 4.2.2. Construct items so the student's task is explicitly described (i.e. what to write about).
     - 4.2.3. Provide students with a sense of time (and effort) to be expended relative to item's value (points).
     - 4.2.4. Do NOT employ optional essay questions.
     - 4.2.5. Try an item yourself by composing, mentally or in writing, a possible response.
     - 4.2.6. Develop an item's scoring key (holistic OR analytic) in advance and employ it with anonymity during grading (see 4.2.7 below).
     - 4.2.7. (Not from Popham but from Stuve). When grading essay questions, **score by item not by student.** This is a process to ensure anonymity and to mask knowledge of scores coming in from auto-graded selected response questions. Avoids "halo", "devil's horns", or pity effects on scoring judgements (and, potentially, item design):
       - 4.2.7.1. Find ways to hide the student's name on instrument as you score.
       - 4.2.7.2. Grade essay questions without knowledge of outcomes from selected response questions.
       - 4.2.7.3. Grade all students' responses to a question before moving to the next question.