

2

Techniques for Testing

Constructing test items and tasks for any type of assessment is a task filled with challenges. Why? Test items are the foundation of tests and the backbone of most assessment instruments.

With her exam specifications firmly in hand, Ms. Wright starts to prepare her midterm exam. In doing so, she does the following to ensure that her students are assessed fairly.

- She balances objective and subjective formats, ensuring that the ones she uses are familiar to students from classroom instruction.
- She derives the weightings in her exam from the focus in curricular outcomes.
- She strives for authentic use of target language through texts and communicative tasks with rich content.
- She has a colleague moderate her test and uses this to check timing, allowing three to four times the teacher's rate for students.
- She pilots her exam via a practice exam with similar rubrics, formats, content, and weighting.
- She plans to use the same basic test for different classes but randomizes questions and answers for objective items, and she writes a parallel prompt for subjective items.
- She gives students a clear idea of the number of points per section and expectations for writing to encourage good time management.

Classifying Test Items and Tasks

Test items can be classified in a number of ways. Some testers categorize items as *selection* and *supply* items. With selection items, a student selects the correct answer from a number of presented options. True/False, matching, and multiple choice are examples of selection items. With supply items, students must supply or construct the correct answer. Examples of supply items include short answer or completion, cloze, gap fill, and essay questions.

Teachers need to be aware of the implications for English language learners in choosing different test formats. The type of response can impact a student's ability to demonstrate what she or he actually knows or can do. For example, a student asked about the plot or sequence of a story might be entirely capable of selecting an appropriate answer. However, the same student with limited English may not be able to supply or create an answer that indicates that they actually understand the content of the story. This simple chart may help you decide what formats to use based on the ability of your students to supply their own answers.

Selection	True/False, Multiple Choice, Matching, Numbering Sequence
Supply	Cloze or Gap-Fill (no responses provided), Essay questions

Subjective or Objective Questions?

Items can also be classified by the way they are scored. Objective test items can be scored based only on following an answer key. Scoring objective items requires no expert judgment, specialist knowledge, or subjectivity on the part of the marker. Scoring subjective items, on the other hand, requires that the marker have knowledge of the content area being tested. Marking a subjective test frequently depends on impression, human judgment, and opinion at the time of the scoring.

In addition to the differences mentioned, subjective/objective item types have these characteristics.

Objective items are usually short answer-closed response items. These types of items test recognition mostly. Although they are usually quick and easy to grade, objective items are generally difficult to write well. In addition, if there are enough of them, they are quite reliable. The great majority of the

workload for teachers who develop these items should take place before the test administration.

Subjective items usually require students to produce longer, more open-ended responses. The emphasis here is on production, as students are generally required to come up with an answer rather than select it from a list of alternatives. Subjective questions are generally easier to write than objective questions, but difficult and time consuming to mark. Objective items have relatively few response options while subjective items are open-ended so a lot of variation is possible in the responses. Reliability of subjective items can sometimes be problematic because these items require human scoring. Issues with inter-rater reliability are sometimes present in subjective-item types. For teachers who routinely use subjective items to test their students, the workload takes place after the test has been administered.

Objective test items are very popular with language teachers and test developers for two reasons. First, these items are easy and quick to mark. Second, they are flexible in that objective test items can be used to test both global and detailed understanding of a text or focus on specific areas of language like grammar and vocabulary. This chapter describes some of the most commonly used objective test items in English language testing, namely multiple choice questions (MCQ), True/False statements (T/F), and the matching format.

Multiple Choice Questions

Multiple choice questions (MCQs) are probably the most commonly used format in professionally developed tests. Teachers all over the world are familiar with the format from their own learning experience, and they understand how it works. Moreover, MCQs are widely used in textbooks as well as on high-profile English language proficiency exams. They are widely used to assess learning at the recall and comprehension levels. Although they are more difficult to write than True/False questions, the job becomes easier with the correct training and a little practice.

MCQs take many forms, but their basic structure is the stem and response. It is the test-taker's task to identify the correct or most appropriate choice. The *stem* is usually written as a question (i.e., *Where did John go?*) or an incomplete statement (i.e., *John _____ to the store*). The *response options* of an MCQ are the choices given to the test-taker. Typically, there are four choices when testing

reading, vocabulary, and grammar, but just three with listening. Only three response options are recommended for listening assessment because students hear the input listening passage only once or twice. Four options put too much load on memory as well as require more reading when we are testing listening. The response options are most commonly expressed as A, B, C, and D. One of these response options is the *key* or correct answer. The others are referred to as *distractors* or incorrect response options. The purpose of distractors is to move students' attention away from the key if they do not know the correct answer, thus determining students' knowledge or skill.

The popularity of MCQs is based on several advantages associated with this format. First, if they are written well, they are very reliable because there is only one answer possible. Second, they can be useful at various educational levels—they are used to assess language at the elementary level and content associated with graduate-level language education. Third, assessment is not affected by test-takers' writing abilities because they are only required to circle the correct response, pencil in a bubble on a sheet, or click on the right answer option. In addition, MCQs are well liked by administrators as being a cost-effective format because they can be scored by computer if the institution has the correct equipment, which makes them quite easy to analyze. Last, students everywhere are familiar with this format.

However, there are some distinct disadvantages to using MCQs. The one most cited by teachers is that MCQs do not lend themselves to the testing of productive language skills or language as communication. Since they are often used to test recognition, teachers forget that they can also be used to assess higher-order thinking skills. Another disadvantage is that MCQs encourage guessing, which can have an effect on exam results. A fourth disadvantage, one that most teachers do not appreciate, is that it is challenging and time consuming to write plausible distractors and produce good items.

Common MCQ Item Violations

In presenting examples of the most common MCQ item violations, we suggest ways to repair them.

- ***Grammatical inconsistency***

A common mistake when developing MCQs is grammatical inconsistency between the stem and the response options. Almost always, the stem and the key are grammatically consistent, but distractors sometimes do not mesh properly with the stem.

Jane spent most of the day at the mall, but she _____ anything.

- A. didn't buy
- B. bought
- C. not buy
- D. many shops

In this item, of course, A is the key or answer. Distractor D is grammatically inconsistent with the other response options as it is a noun clause while the others are all verb forms. To fix this item, distractor D should be changed to a verb form like *buying*.

- ***Extraneous cues or clues***

Cueing can occur in two places on a test: within an item or within the test. An extraneous clue that occurs within the test is one where students can find the answer to a question somewhere else on the test paper in another section of the test. Consider the following cueing violation within an item:

After I've had a big lunch, I only want an _____ for dinner.

- A. pot of soup
- B. apple
- C. big steak
- D. candy bar

The key here is B because it is the only distractor that takes the article *an* in the stem. In this item, a student only needs to know the grammatical rules concerning *a* and *an* to figure out the correct response. To fix this item, consider putting *a/an* in the stem.

- **3 for 1 split**

This item violation occurs when three distractors are parallel and one is not. It is sometimes called *odd man out*. This item violation varies in the degree of seriousness. It is a serious violation if the unparallel option is the key.

The company was in desperate need for more workers, so they _____ an expensive ad in the newspaper.

- A. placing
- B. to place
- C. placement
- D. placed

In this item, D is the key. The 3 for 1 split is three verb forms (A, B, and D) and 1 noun (C). (As 3 for 1 splits go, this is not a terrible one because the key is not the odd man out, but often the odd man is the key.)

- **Impure items**

Impure items are those that test more than one thing.

I didn't see _____.

- A. had the students gone
- B. the students had gone
- C. have the students gone
- D. that students have gone

This item tests both verb tense and word order. Remember that good items should only test one concept or point.

- **Apples and oranges**

An apples-and-oranges violation is one where two response options have no relation to the other two. This is often referred to as a 2 for 2 split. There are instances where 2 for 2 splits are acceptable (i.e., the case of opposites or affirmative/negative).

According to the reading passage, people use mobile phones _____.

- A. frequently
- B. seldom
- C. in their cars
- D. for emergency purposes

Distractors A and B are adverbs and C and D are prepositional phrases. This item would be better if all the response options were either adverbs or prepositional phrases and only one was clearly the key. Without reference to the reading passage, all four options are potential keys.

- **Subsuming response options**

In this item violation, the intended answer and a very good distractor could both be correct. Consider this sample listening item.

Mary: We need to buy some new lawn furniture so we can sit on the patio.

Steve: Okay. I'll go to the mall tonight. Any special kind you're looking for?

Mary: Something cheap and comfortable. Just don't get anything made of metal, please.

What will Steve buy?

- A. outdoor furniture
- B. comfortable chairs
- C. steel furnishings

Although the key is B, it can be argued that A is correct on a higher level because *comfortable chairs* are part of *outdoor furniture*. It is a sign of a poor test item when two response options can be considered correct.

- **Unparallel options**

This item violation occurs when the response options are not parallel either in length or in grammatical consistency.

How do students in the suburbs usually travel to school?

- A. by bus
- B. they go by taxi
- C. most of them either take the bus or get a lift from their parents
- D. walk

This item is not parallel in either grammar or in length. Distractor C is the key, and it is by far the longest response option. All response options should be the same part of speech and approximately the same length.

- *Gender bias in language*

Particular care should be taken when using vocabulary that relates to gender. For example:

My cousin works as a male nurse _____ General Hospital.

- A. over
- B. on
- C. from
- D. at

The term *nurse* refers to both men and women who work in this profession. We don't say "female teacher," so why the need to say "male nurse"?

- *Sensitivity*

Materials that have a negative emotional impact on students should be avoided. Similarly, test item content that could be upsetting or shocking to students is not advisable.

The _____ disaster of 2004 was the December 26th tsunami that struck southeast Asia.

- A. most worst
- B. baddest
- C. very bad
- D. worst

For a group of ESL students from Mexico, this item is not a problem. However, for students who might be from a country hit by the tsunami, this item could be very upsetting. Test-taking is a very scary experience for most students, so it is important that the content they encounter on tests not add to that fear.

- *Double answer or key*

This item violation is the most commonly made among teachers. It occurs when more than one response option is correct.

The teacher waited in her office until her students _____.

- A. came
- B. would come
- C. come
- D. had come

Many would argue that both A and D are correct responses.

- ***No answer***

This is a common item violation made by teachers. It occurs when the author of the test item forgets to include the key among the list of response options. This most often occurs when the item has undergone various revisions and rewrites.

This is the restaurant _____ I told you about yesterday.

- A. what
- B. where
- C. why
- D. how

- ***Giveaway distractors***

This violation occurs when test-takers are able to improve their scores by eliminating absurd or giveaway distractors.

According to the text, the author of the article comes from _____.

- A. Dubai
- B. France
- C. Buenos Aires
- D. Disneyland

Students can easily eliminate D so it is not an effective distractor. It adds nothing to this question.

Giveaway!

Tips for Writing Good Multiple Choice Questions

Multiple choice questions are the hardest type of objective question to write for classroom teachers. Teachers should keep the following guiding principles in mind when writing MCQs.

- **The question or task should be clear from the stem of the MCQ.**

Write MCQs that test only one concept.

- **Take background knowledge into account.**

The selection of the correct or best answer should involve interpretation of the passage/stem, not merely the activation of background knowledge.

- **Provide as much context as possible.**

The MCQ format is often criticized for its lack of authenticity. Whenever possible, set items in context.

- **Keep sensitivity and fairness issues in mind.**

It is important to write items that do not unfairly advantage or disadvantage certain groups of students. It is also crucial to avoid item content that could be offensive or upsetting.

- **Standardize the number of response options.**

The optimum number of response options for foreign/second language testing is *four* for grammar, vocabulary, and reading, and *three* for listening. However, some testers feel that three response options are acceptable for classroom or progress testing. Although there is no psychometric advantage to having a uniform number (Frery, 1995), it is better to be consistent within a test.

- **One response option should be an unambiguous correct or best answer.**

The three remaining options function as distractors. Distractors should attract students who are unsure of the answer.

- **All response options should be similar in length and level of difficulty.**

Response options should be consistent with the stem and parallel in length and grammar with each other. Some teachers inadvertently make the answer longer and more detailed than the distractors

simply because they add information to make it unambiguously correct. Testwise students often spot this practice.

- **Avoid using distractors like *none of the above*, or *a, b, and sometimes c*, but never *d* options.**

Avoid these because they test more than one thing at a time. The distractor *all of the above* is still acceptable in certain cases, but its use remains controversial. Those who are not in favor of it state that the recognition of two right options identifies it as the answer. Additionally, most teachers use it as the correct answer to almost every item containing it as a response option, and the test-wise student soon figures this out. Therefore, proponents of *all of the above* recommend using it as an incorrect answer as many times as a correct answer.

- **Correct answers should appear equally in all positions.**

Randomly assign correct answers. Don't unconsciously introduce a pattern into the test that will help the students who are guessing or who do not know the answer to get the correct answer. Although it may make grading easier, it is a threat to reliability. Don't neglect placing the answer in the A position. Research has shown that this is the most neglected position because teachers want students to read through all response options before responding to the question, so teachers unconsciously place the actual answer further down in the list of response options. One way to randomize answers is to alphabetize them. By doing so, the correct answer will automatically vary in the A, B, C, or D position.

- **Move recurring information in response options to the stem.**

If the same words appear in all response options, take these words out of the response options and put them in the stem.

- **Avoid writing absurd or giveaway distractors.**

Do not waste space by including funny or implausible distractors in your items. All distractors should appear for a valid pedagogical reason.

- **Avoid extraneous clues.**

Avoid unintentional grammatical, phonological, or morphological clues that assist students in answering an item without having the requisite knowledge or skill being tested.

- **Make the stem positive.**

Writing the stem in the affirmative tends to make the question more understandable. Introducing negatives increases the difficulty and discrimination of the question. However, if you must make it negative, place the negative near the end of the statement (i.e., *Which of the following is NOT.* or *All of the following are _____ except. . . .*).

- **Make sure all questions are independent of one another.**

Avoid sequential items where the successful completion of one question presupposes a correct answer to the preceding question. This presents students with a double jeopardy situation. If they answer the first question incorrectly, they automatically miss the second question, thereby penalizing them twice.

- **Use statistics to help you understand your MCQs.**

Statistics like item analysis can assist you in your decisions about items. Use statistics to help you decide whether to accept, discard, or revise MCQs. (See Chapter 9 for information on item analysis.)

True/False Format

True/False questions are second only to multiple choice questions in frequency of use in professionally produced tests and perhaps one of the most popular formats for teacher-produced tests. Basically, they are a specialized form of the MCQ format in which there are only two possible alternatives and where students must classify their answers into one of two response categories. The common response categories are: *True/False*, *yes/no*, *correct/incorrect*, *right/wrong*, or *fact/opinion*. Because True/False is the most common response category, these questions are generally referred to as True/False questions.

True/False questions are typically written as statements, and the students' task is to decide whether they are True or False. They are attractive to many test developers because they offer several advantages. First, when you use this question type, you can test large amounts of content. Additionally, because True/False questions are shorter than most other item types, they typically require less time for students to respond to them. Consequently, more items can be incorporated into tests than is possible with other item types, which

increases reliability. Another big advantage is that scoring is quick and reliable and can be accomplished efficiently and accurately.

Despite their advantages, True/False questions have several disadvantages. One is that there is a 50 percent guessing factor that the choice will be correct. With only two possible answers, there is always the danger that guessing may distort or inflate the final mark. To overcome this disadvantage, it is recommended that teachers use a third response category called Not Given or Not Enough Information. By doing so, the guessing factor goes from 50 percent to a more acceptable 33.3 percent. Yet another way to alleviate this problem is to ask students to correct false statements or to find statements in the text that support either a true or a false response. These two methods increase the diagnostic value of True/False questions. A second disadvantage of True/False items is that for them to be reliable, you need to include a sufficient number of them on the test.

Teacup Dogs

Teacup dogs are the latest must-have Hollywood fashion accessory. They get their name because these tiny dogs fit rather nicely into a teacup. The latest Hollywood trend weighs less than three pounds. You only need to flip through the pages of celebrity magazines to see these adorable dogs peeking out of designer handbags on the streets of New York and Hollywood. A trend originally started by Paris Hilton's Chihuahua, Tinkerbelle, now other celebrities proudly sport tiny dogs as fashion accessories.

Animal cruelty activists, however, state that this fashion statement has gone too far. They point out that due to increased pressure from buyers, breeders are attempting to downsize these dogs still further. Because of these breeding practices, these darling dogs are often unhealthy creatures. Veterinarians point out that teacup dog owners have to cope with high medical bills as teacup dogs are prone to genetic diseases like water on the brain, heart problems, knee problems, and dental disease. Even more sadly, this wide range of medical problems causes them to have a shorter life span than normal dogs.

So are teacup dogs the latest fashion statement or an example of cruelty to animals? You decide.

Answer the following questions.

Poor T/F question: Teacup dogs weigh less than three pounds.

Better T/F question: Teacup dogs are much smaller than normal dogs.

The first question is more or less a verbatim matching from the text. The latter option is better as it paraphrases the information.

Poor T/F question: Teacup dogs are only found in New York and Hollywood.

The word *only* in the first option is an absoluteness clue to this sentence being false as you only need one exception (in this case one other teacup dog in another city) to make it false.

Poor T/F question: The author of this article is probably a proponent of teacup dogs.

Better T/F question: The author of this article probably likes teacup dogs.

The word *proponent* in the first option makes this question very difficult. An easier synonym like *supporter* is less difficult but still assesses the same principle. Simplifying vocabulary ensures that you are testing a target concept, not the confounding vocabulary in the question itself. This attention to potentially problematic vocabulary is especially important to K–12 audiences.

Tips for Writing Good True/False Questions

The following tips can help you write effective True/False questions.

- **Write items that test meaning rather than trivial detail.**
True/False items are said to test gist or intensive understanding very well.
- **Questions should be written at a lower level of language difficulty than the text.**
This is important because you want to ensure that comprehension is based on understanding of the text and not understanding of the question itself. (This is important for lower-proficiency learners, especially K–12 learners.)
- **Consider the effects of background knowledge.**
Successful completion of True/False/Not Given items should depend on the students' reading of the text, not on background knowledge.

- **Questions should appear in the same order as the answers appear in the text.**

By mixing the order of True/False questions in reading and listening, you increase the difficulty of these questions significantly.

- **Make sure you paraphrase questions in simple, clear language.**

It is better to paraphrase questions rather than take them verbatim from the text. The latter only requires students to locate the relevant statements in the text to be able to answer them. Paraphrase by using vocabulary and grammar from course materials. The language of the question should be simple and clear yet have sufficient information to allow its truthfulness to be judged.

- **Avoid absoluteness clues.**

Do not use specific determiners like *all*, *none*, *always*, and *never*. Questions with these determiners are easy because the answer is most always false. Similarly, avoid determiners like *sometimes* and *some* as they tend to appear in statements that are true.

- **Focus each item on a single idea from the text.**

Items that require students to deal with the possible truth or falsity of two or more ideas at once increase the difficulty of the question substantially.

- **Avoid answer patterns.**

Don't be tempted to write questions with a specific answer pattern like TTFFTTFFTT to facilitate your grading. Students will soon catch on to these tactics. Answer patterns in your questions should not be discernable.

- **Include enough questions.**

True/False questions are a reliable way of testing students' comprehension if there are enough items. It is recommended that teachers include a minimum of seven to ten questions on their tests when using this format.

- **Add a third option to decrease the guessing factor.**

The True/False format has a high guessing factor. By adding a third response category (Not Given or Not Enough Information), teachers can decrease this guessing factor. It should be noted that the NG/NI option is appropriate for students at the intermediate level and

higher and should not be used in the assessment of listening comprehension. In reading comprehension, students have unlimited opportunities to go back to the text to determine if content is not given, but in listening comprehension, students hear the source text only once or twice. Including an NG option would tax students' memory.

- **Have students circle T, F, or N on the test paper or answer sheet.**

By doing so, you will avoid getting those Ts that suspiciously look like Fs. This will substantially facilitate your marking.

Matching Format

Another common objective format is matching. Matching is an extended form of MCQ that draws on the student's ability to make connections among ideas, vocabulary, and structure. Matching questions present the students with two columns of information. Students must find the matches between the two columns. Items in the left-hand column are called *premises* or *stems*, and the items in the right-hand column are called *options*. The advantage of matching questions over MCQs is that the student has more distractors per item. Additionally, writing items in the matching format is somewhat easier for teachers than either MCQs or True/False/Not Given.

Consider this example assessing proverbs.

Poor Matching Question Set:

- | | |
|----------------------|--------------------------|
| 1. better late | A. wear it |
| 2. if the shoe fits, | B. keeps the doctor away |
| 3. an apple a day | C. is not gold |
| 4. the early bird | D. catches the worm |
| 5. all that glitters | E. than never |

Both the options and the premises have an equal number of choices. Therefore if students miss one, they miss at least two automatically. Similarly, if they get four correct, they get the last one correct by default. When formatting matching questions, it is better to draw a line before the numbers so that students can write "the letter of the correct answer in the space provided." With no blanks, in order to answer students will have to draw lines from option to premise, making grading very difficult.