10 Choosing and devising test tasks

- Introduction
- Choosing test tasks
- Guidelines for making open-ended test tasks
- Guidelines for making closed-ended test tasks
- Summary

Preview questions

1. What are the advantages of tests in comparison to the other methods of collecting information for evaluation that we have discussed in this book? What are their disadvantages?
2. Do you use (or have you used) multiple-choice questions to test your students' language skills? What did you find difficult about making up such tests? What did you find useful about them?
3. Do you use essay-type (or open-ended type) questions in your tests? When and why do you use this format? What are the easy and the difficult parts to making up and using such test formats?
4. If you have ever taken multiple-choice tests, what did you personally like and dislike about them? Did you feel that your performance was a fair reflection of what you knew? If not, why not?
5. When you have had to do essay-type tests, what did you personally find difficult about them? Did you have to study in any particular way for such a test in comparison with multiple-choice tests? Is one way of studying better than another?
6. What distinguishes authentic language use from nonauthentic language? Suggest some examples of authentic language use and some ways of testing proficiency in using language in these ways.
7. Are there some authentic language tasks that could be tested validly using multiple-choice task formats? Name them.

Introduction

In this chapter, we talk about how to choose among the three general test task types reviewed in Chapter 9. We also present guidelines for devising closed-ended and open-ended test tasks; we do not discuss limited-response formats because the guidelines presented for the open- and closed-ended formats can be adapted for this purpose. These guidelines are part of the larger process of devising valid tests that are compatible with instructional objectives and, specifically, with the focus, range, and standards of performance specified or included in your objectives.

Choosing test tasks

Choosing the type or types of tasks to include in a language test depends on a combination of factors:

1. Instructional objectives
2. The students' level of proficiency
3. Instructional activities
4. Available testing resources

What follows are general suggestions to assist you in the selection of test tasks.

Instructional objectives

Clearly, the most important factor to consider when choosing which type of test task to use is your objectives. Choose tasks that focus on the same kinds of language skills described in the objectives as well as the range and standards of performance expected of the students. Closed-ended tasks permit assessment of comprehension skills in both reading and listening, but they do not lend themselves to directly assessing production skills: speaking or writing. This is to say, one's ability to perform on a closed-ended test task does not necessarily mean that the individual would be able to produce the corresponding language in an open-ended task. Also related to language objectives, closed-ended tasks permit the examiner to assess specific language skills – this follows from the fact that the responses permitted by closed-ended tasks are controlled totally by the examiner. In comparison, limited-response and open-ended response tasks do not control the students' specific responses -- students can often find ways of responding to test items that are different from what was intended by the examiner.

The range of language skills elicited by a closed-ended task is strictly under the control of test makers: they can include as broad or as narrow a range of language skills as desired. Moreover, closed-ended tasks force the test taker to respond to test items in specific ways so that the examiner can examine a specified range of skills. In comparison, test makers cannot control the range of language skills elicited by open-ended tasks. In principle, an open-ended task could elicit a very broad range of skills. In practice,

Cambridge University Press
978-0-521-56681-0 - Classroom-Based Evaluation in Second Language Education
Fred Genesee and John A. Upshur
Excerpt
More information

however, learners may, and often do, limit their responses to those skills they have some confidence in. Thus, weaker students might produce a much more restricted range of language in response to a composition, for example, than more proficient students. Their performance may be nevertheless linguistically correct. If students do not use certain linguistic items or structures in an open-ended task, it is not possible to tell whether they do not know them or whether they simply chose not to use them. Thus, on the one hand, open-ended tasks can yield very rich samples of language and, on the other hand, may yield restricted samples because students choose not to use as broad a range of language as hoped for or because they avoid using language they do not have complete control over.

The same issue arises when considering the selection of open- versus closed-ended test tasks from the point of view of standards of performance. To the extent that open-ended tasks permit students to not use language that might be of interest to the examiner, then the examiner may not be able to assess the students' performance thoroughly with respect to certain standards of performance. Students can often find ingenious ways of avoiding language they do not know or know only poorly. In comparison, closed-ended tasks force students to respond to a limited range of alternatives that can be selected carefully to represent the standards of performance of interest to the examiner. At the same time, closed-ended tasks assess only recognition skills and, therefore, may not fully capture students' ability to actively use language according to these standards.

Level of proficiency

Closed-ended and limited-response tasks can be particularly suitable for assessing the language skills of beginning level second language learners. This does not mean that closed-ended and limited-response formats cannot be used for intermediate or advanced level students. Whether such tasks are suitable for more advanced students will depend upon the exact content of the item, not on the response characteristics per se. Open-ended tasks, in comparison, can be particularly suitable for assessing more advanced students. If different task types are used in a single test, it is generally desirable to start off with closed-ended tasks in order to put students at ease and to include limited- or open-ended response items later once the students have warmed up.

Instructional activities

Test tasks should be chosen by taking into account the kinds of instructional activities the learners have been exposed to. This ensures that students are familiar with and, therefore, understand the response demands of the task. It is unfair, for example, to use open-ended response tasks with learners who have been exposed to only closed-ended kinds of learning activities. Therefore, test tasks should always be chosen that are well understood by students, either by virtue of their classroom experiences with similar tasks during instruction or by virtue of clear instructions in the test.

Testing resources

Finally, test tasks should be practical given the resources available. An important resource to consider is time, both for administering the test and for scoring it. In general, open-ended test tasks take much longer to score than closed-ended or limited-response tasks. Either type of task can take a brief or a long time to administer, depending on the content of the test. The physical resources for testing are also important. Individual testing that requires private, quiet space (e.g., oral interviews) is impractical if the examiner does not have a separate area for conducting the interviews. Such a task might also be impractical if the examiner does not have the human resources to supervise other students who are not being tested.

Task 1

Identify a language skill to be tested (e.g., listening comprehension), and then brainstorm alternative open-ended and closed-ended tasks to test it. Discuss the merits of each alternative.

Guidelines for making open-ended test tasks

Introduction

As we just noted, in contrast to closed-ended test tasks, open-ended tasks do not control in a precise way the specific responses to be made by the test taker. Students are relatively free to respond in whatever way they choose. For example, in an oral interview, each test taker can respond to the interviewer's questions in a unique way, using different language structures, vocabulary, and so on. These kinds of tasks are particularly suitable for assessing language proficiency in many authentic situations that call for interactive, dynamic language use – such as a conversation between people or an encounter with a salesclerk in a store or a stranger on the street. Although many such situations are open-ended, they are always structured in particular ways. For example, a conversation has a certain structure or organization to it, although the topics of conversation may vary considerably; or a job interview also has a specific structure, and the interviewee can imagine beforehand the kinds of questions that may be asked. Assessing language proficiency in ways that resemble the actual situations and tasks for which language learning is taking place is sometimes called performance-based assessment.

At the same time, it is important to realize that not all authentic language tasks are open-ended; for example, filling out application forms and buying bus tickets, stamps, or gasoline are quite formulaic. Not all authentic language use involves oral communication; reading and writing are also characteristic of much authentic language use. Even taking multiple-choice tests can be an authentic language task for second language learners who are studying in schools in which the second language is the medium of instruction. Language performance in school often, although not always, calls for the ability to take tests, and preparing second language learners for such activities is common in many English language universities that enroll large numbers of nonnative speakers.

Because they are less structured than closed-ended tasks, open-ended tasks are often used to assess the skills of advanced level learners. In contrast, beginning level learners often need the structure imposed by closed-ended and limited-response tasks; oral tests for beginners, for example, often include such activities as picture naming and question answering. However, multiple-choice tasks, although well suited for testing beginning level learners, can nevertheless be demanding if care is not taken to avoid unnecessary complications.

Open-ended test tasks are suitable for testing speaking and writing skills because they require language production. They tend to be used to assess higher order skills, such as discourse and sociolinguistic skills in particular that cannot be elicited easily using closed-ended or limited-response test tasks. In fact, open-ended tasks call for a variety of language skills. For example, a written composition requires spelling, vocabulary, and grammar skills in addition to discourse and sociolinguistic skills. Thus, it is possible to score open-ended tasks for different language skills.

A great deal of judgment is called for when scoring open-ended tasks because each student's response can be different from other students' responses but no less correct. Consequently, scoring open-ended tasks is much more demanding and requires much more thought than scoring closed-ended tests. Moreover, if open-ended test tasks are used to assess language proficiency in authentic situations, then judgments of appropriateness, effectiveness, and correctness are often called for since these are important standards for assessing language use in situations in which language is normally used. Indeed, normally, correct use of language is not an end in itself but a means for negotiating social relations, transacting business, or achieving other goals. For instance, teaching assistants at the university level use language in order to help their students understand course material; computer salespersons use language to sell computers; and doctors use language to understand the source of their patients' medical problems. Language is vital for these people in the overall performance of their duties and jobs. Evidence of the success of their language skills lies in how well they perform these duties, not simply in how correctly they use language. Even when there is no face-to-face interaction, language use normally involves some form of interaction – someone who is listening or reading someone else's message. Even in these cases, appropriateness and effectiveness of communication can be important standards for assessment.

Because the specific responses to be made by test takers in open-ended tasks are not controlled in any precise way, devising such tasks does not require the same precision or technical care as closed-ended tasks, although they may require some ingenuity to ensure that the test task resembles the kinds of situations in which the learners will ultimately use their second language. Open-ended tests are different from closed-ended tests in that they usually consist of only one item (e.g., write a 250-word essay on a topic of your choice), although this is not always the case. In contrast, tests made up of closed-ended tasks generally include a number of items.

The guidelines in this section take the form of *general questions* you can ask about open-ended tasks rather than specific technical suggestions of the type provided for closed-ended test tasks.

General questions

When constructing open-ended test tasks, it is useful to ask the following general questions:

1. Is the task appropriate with respect to instructional objectives and instructional activities?
2. Is the task understandable with respect to expected performance and assessment standards?
3. Is the task feasible with respect to topic, level of difficulty, and time?

Each of these general questions contains a number of specific questions. These are summarized in Table 1 and discussed next.

APPROPRIATE

When selecting an open-ended test task, follow the same general process used when choosing closed-ended test tasks; that is to say, it is important to select a task that is valid with respect to your instructional objectives. More

Table 1. Guidelines for devising open-ended tests

Appropriateness

1. Can the task elicit the kinds of language skills identified in the instructional objectives?
2. Can the task elicit the range of language skills identified in the instructional objectives?
3. Do the language skills elicited by the task lend themselves to assessing the students' performance according to the standards expected of them?
4. Does the task reflect the actual performance demands of the situations in which the second language will ultimately be used?
5. Are the students prepared for the task?
6. Is this task workable with the students?

Understandability

1. Have the task demands of the test been made explicit and clear?
2. Have the standards of performance and evaluation been made explicit to the students?

Feasibility

1. Will the topic of the task elicit the kinds of language skills you want to examine?
2. Is the topic of interest to the students?
3. Is the topic biased?
4. Is the specific form of the task of appropriate difficulty?
5. Is there enough time for the students to complete the task? Conversely, has so much time been allotted that the test no longer reflects normal time constraints?

specifically, choose a task that reflects (1) the same linguistic focus, (2) the range of performance specified by the objectives, and (3) the standards of performance expected of the students. When using open-ended test tasks to assess language proficiency in authentic situations, try to select tasks that approximate the actual situations in which the students will use their second language skills as much as possible so that you elicit these skills and so that you can, in turn, make accurate predictions of your students' language performance. Most situations in which language is normally used are interactive, dynamic, and purposive in ways that extend beyond simply using language correctly. If your instructional objectives aim for proficient use of language in such situations, then an appropriate test task should also include these qualities. In addition, one should ensure that the full range of performance standards is part of the scoring system, including measures of how accurate, appropriate, and effective the students' performance is.

In some cases, special efforts need to be taken to simulate the performance demands of authentic situations in which the second language will ultimately be used because the target situations are different from those in the classroom. For example, in courses for people who are learning ESL for business purposes, the actual target situations are not likely to be part of the second language classroom. Teachers will need to exercise some ingenuity to simulate in their classroom the actual situations in which the language will be used if they want their assessment to reflect authentic language use. In contrast, when teaching ESL for academic purposes, the target situations in which English will ultimately be used can probably be found quite easily in the second language classroom itself. Much less effort is needed in these latter instances to create a testing situation that reflects the task demands of the target situation. The more closely your test task simulates the actual conditions in which the second language will be used, the greater the predictive validity of your test results.

When selecting appropriate test tasks, take into account the instructional activities that have been used in class. Students may not be able to demonstrate the full extent of their proficiency if a test task is selected that they have not seen before because the task demands may not be clear to them. For example, using role play for the first time as a test may not work with your students because they do not know what is called for, or they may simply be too self-conscious. Using the same kinds of tasks as have been used as instructional activities ensures that your students are familiar with the task demands.

At the same time, using exactly the same activities that were used in class will not tell you whether students can use their new language skills in new but related situations. Surely an indication of language proficiency is the ability to use language in different situations. This is particularly relevant when testing students at advanced levels of proficiency where generalizability would be expected. Judgment is called for when selecting test tasks that are different from but related to the activities you have used in class. You cannot know with any certainty whether your students can handle a new situation until you have tried it out.

Thus, when devising test tasks so that they are appropriate, there are a number of specific questions to ask:

1. Can the task elicit the kinds of language skills identified in the instructional objectives?
2. Can the task elicit the range of language skills identified in the instructional objectives?
3. Do the language skills elicited by the task lend themselves to assessing the students' performance according to the standards expected of them?

4. Does the task reflect the actual performance demands of the situations in which the second language will ultimately be used?
5. Are the students prepared for the task?
6. Is this task workable with the students?

UNDERSTANDABLE

Because open-ended test tasks allow wide variation in responding, test takers must understand what is expected of them. Test tasks that are not well understood become puzzles that require the student to guess what the examiner wants. Test takers who do not know what is expected of them might give wrong or inappropriate responses because of misunderstanding and not because of lack of language proficiency. Test instructions should be simple, straightforward, and unambiguous. Students should also have some specific indications of what counts in judging their performance. In a written composition, for example, does spelling count? Is originality important? In an oral interview, what exactly will be scored: pronunciation, grammar, the organization of their responses? What weight will different scoring components be given? Students need to be well informed in order to decide how to spend their time and energy during the test. The standards of performance that will be used to judge language performance should be made clear to the students prior to testing. Deciding on a scoring scheme after the test has been given or informing students of scoring criteria after the test is unfair.

Specific questions that can be asked about understandability are:

1. Have the task demands of the test been made explicit and clear to the students?
2. Have the standards of performance and evaluation been made explicit to the students?

FEASIBLE

Having chosen a certain open-ended test task, you must decide whether the task is feasible. There are at least three aspects of test tasks to examine from the point of view of feasibility: (1) task topic, (2) task difficulty, and (3) the time allotted to perform the task.

- Topic Will or can the topic you have chosen elicit the kinds of language skills you are interested in? Sometimes topics that instructors think will work, do not. You may have to try them out beforehand with other students in order to determine this. Or even trying to do the task yourself can give you a general indication of the feasibility of the topic.

Is the topic realistic and authentic? Using topics that students do not regard as authentic will reduce the legitimacy of the test in your students' eyes and certainly will not elicit authentic language performance. Is the topic of interest to the test takers? If it is not, the test takers will not be motivated to respond seriously or enthusiastically. Interest and motivation are particularly important in open-ended tasks because the test takers are free to respond as much or as little as they want. Does the topic favor or disfavor individuals or subgroups of learners for reasons that have nothing to do with the course? That is to say, is there unfair bias in the topic? For example, is the topic culturally offensive to certain students? Do some of them have additional experience with the topic, such as a science topic that would allow students with a science background to perform better than students who do not have such a background?

From the examiner's point of view, will you be able to get the students to respond to the topic? Will the language samples produced in response to this task allow you to form a realistic picture of the student's ability with respect to the objectives you are testing? And can the language samples elicited by this topic be scored appropriately?

- Difficulty Is the task of appropriate difficulty, or is it so difficult that students will be unable to demonstrate the language skills they have acquired? Conversely, is it so easy that all students will find it trivial or unchallenging?

From the examiner's point of view, is the exercise so easy that scores will fail to distinguish those students who have made more progress from those who have not progressed as much? Is the task so difficult or complex that the examiner will find it difficult to determine what anyone has learned?

- Time Is there enough time for students to perform the task? On the one hand, students who are not given enough time will not be able to demonstrate their full achievement. On the other hand, students who are given too much time to do a test can treat it like a puzzle rather than an actual language task. So-called speeded or time-constrained tests are appropriate sometimes – namely, when the language skill they are testing is usually performed with time constraints; for example, an impromptu oral report or conversation should have time constraints but writing academic assignments probably should not. Speeded tests are usually used with material that is so easy that, given enough time, all test takers would be expected to respond correctly. Consequently, the test takers are being examined on their speed of performance rather than their skill or knowledge alone.

In contrast, a *power* test is one that allows enough time for nearly all test takers to complete it, but the material being examined is of sufficient difficulty that not all test takers are expected to get every item correct. Thus, power tests examine maximum level of skill or knowledge without time constraints. Test performance under speeded conditions is not usually the best indicator of maximum performance capabilities. Whether a speed or power test is appropriate will depend on your objectives. (We discuss time for testing further in Chapter 11.)

Specific questions to ask when considering the feasibility of test tasks include:

1. Will the topic of the task elicit the kinds of language skills you want to examine?
2. Is the topic of interest to the students?
3. Is the topic biased?
4. Is the specific form of the task of appropriate difficulty?
5. Is there enough time for the students to complete the task? Conversely, has so much time been allotted that the test no longer reflects normal time constraints for performance of such tasks?

Task 2

Compare the kind of information provided by a written test with that provided by students' journals or a writing conference. Discuss the uses and limitations of each method.

Guidelines for making closed-ended test tasks

Introduction

Closed-ended response tasks are suitable for testing skills involved in reading and listening because they involve comprehension skills. They do not require the test taker to produce or generate a response. Closed-ended response tasks can be particularly suitable for beginning level learners precisely because they do not require language production and because they are highly structured. Their use is not restricted to beginners, of course, and they can be made as complex as desired depending on the particular nature of the task and its content.

Most closed-ended test tasks are some form of what is commonly known as multiple-choice questions, although there are some variations that are not. Matching tasks in which the test taker must match one set of items, such as specific words, to another set, such as different "parts of speech" or grammatical terms, are an example. However, even this format can be conceived of as multiple-choice in that the grammatical items constitute a set of multiple-choice answers, only one of which is correct as a descriptor of each word. Multiple-choice question formats include a stem, or prompt, and alternative responses. The stem is, in effect, the question. The alternatives that are not correct are called *distractors.*

Closed-ended test tasks attempt to control in precise ways the particular response required to perform the task. Thus, they are especially useful for assessing particular aspects of language, such as certain grammar rules, functions, and vocabulary. A great deal of care is called for in making up these tasks in order to avoid ambiguous or misleading items that are confusing to the test taker and produce answers that are meaningless to the examiner. Thus, closed-ended tasks, and multiple-choice questions in particular, are difficult to construct. However, scoring is simply a matter of checking whether the correct alternative was chosen.

What follows are general guidelines for constructing multiple-choice types of closed-ended test tasks. These guidelines are summarized in Table 2. We present guidelines for preparing stems and response alternatives. Bear in mind that the guidelines presented for open-ended test tasks are also

Table 2. Checklist for devising closed-ended tests

The stem

1. Is the stem simple and concise?
2. Are there unnecessary double negatives or other complex wordings in the stem?
3. Does the stem assess what it is supposed to?
4. Are there inadvertent cues to the right answer?
5. Is the stem a verbatim repetition of material taught in class? If so, is this desirable?

The response alternatives

1. Are the distractors of the same grammatical and semantic class as the correct response?
2. Are the response alternatives grammatically compatible with the stem?
3. Are all the alternatives equally attractive?
4. Are the distractors informative?
5. Are the alternatives equally difficult, complex, and long?
6. Is there more than one correct alternative?
7. Does the wording of the alternatives match the stem?
8. Can the correct response be derived from common knowledge?
9. Are the alternatives suitably simple?
10. Can the answer to any items be derived from other items?
11. Do any of the alternatives refer to other items?

relevant when devising closed-ended tasks. In other words, closed-ended tasks, like open-ended tasks, should be appropriate, understandable, and feasible.

The stem

The stem, or prompt, in a multiple-choice task can be linguistic or nonlinguistic in nature. Nonlinguistic stems consist of pictures or realia (i.e., real objects). Linguistic stems can consist of single words, phrases, sentences, written text, oral passages, or discourse.

1. The stem should be presented in a simple, concise form so that the task or problem posed by the item is clear and unambiguous. In other words, it should be clear to the test taker what is called for after reading or hearing the stem. For example:

   The following item is ambiguous because it leads to more than one possible correct answer:

   - She watched her carefully _______ her coat on.
     a. put*
     b. puts
     c. to put
     d. while putting*

2. In most cases, it is advisable to avoid using negatively worded stems since they make extra and often unnecessary demands on the test taker. For example:

   - Which of the following is not true?

3. Make sure the stem is testing what it is supposed to be testing. In particular, make sure that the point that is being tested is the only source of difficulty in the stem. Otherwise, the task will demand more than you want to test. For example, if the item is testing vocabulary, make sure that any language used to provide context for the target item is familiar and comprehensible to the student; otherwise you might be testing knowledge of more than one vocabulary item. For example:

   - Jane *donated* a candelabrum to the charity bazaar.
     a. gave*
     b. sold
     c. sent
     d. wore

   In this stem, the target word is *donated,* but the use of the words *candelabrum* and *bazaar* may confuse and mislead students because they are not familiar with them.

Or if it is a sentence comprehension item, for example, make sure that obscure vocabulary that could impede comprehension is avoided, that is, unless you want to see whether the students can infer the meaning of unknown words from context. Obscure vocabulary in a sentence comprehension item turns the task from sentence comprehension to vocabulary knowledge.

4. Avoid stems that inadvertently give clues to the right answer for unimportant or uninteresting reasons. For example:

   - Charlie is always late for school, so his mother is going to buy an _______ clock for him.

     a. ring
     b. alarm*
     c. morning
     d. bell

   The correct answer in this case (alarm) is cued by the article *an* in addition to the meaning; *an* is always followed by a word that begins with a vowel, and *alarm* is the only alternative that begins with a vowel.

5. It is often advisable to avoid making the stem identical to material that has been taught or used in class in order to avoid correct responding on the basis of memory alone. There may be exceptions to this, such as testing for comprehension of idiomatic expressions.

The response alternatives

Like the stem, the alternative responses in a multiple-choice item can be linguistic or nonlinguistic in nature. The latter can consist of pictures of realia. For example, the examiner says a word, and the students must select from among four alternative pictures the one that corresponds to the word. More commonly, alternative responses are expressed in linguistic form. For example, the examiner says a word, and the students select from among four spoken alternative words the one that is a synonym of the target.

1. Distractors should belong to the same general grammatical or semantic category as the correct response. In other words, avoid distractors that are different from the correct alternative in structural or semantic terms. If the test taker has a general idea of the type of response called for, such differences might give inadvertent clues that certain responses are wrong.

   For example, if the stem consists of a sentence with a word missing (i.e., cloze format), then all alternative responses should belong to the grammatical category needed to fill in the blank. For example:

- She walked ______ up the steps to the library.

  a. weak
  b. slowly*
  c. try
  d. wisdom

In this case, the test taker might recognize that an adverb is called for. Since *slowly* is the only adverb among the alternatives, it would be selected regardless of its meaning. This would be an appropriate set of distractors if you want to test the test takers' understanding that an adverb, and not some other part of speech, is called for in this gap.

In reading comprehension tasks, the distractors should refer to the text in some way. Alternatives that are totally unrelated to the stem can be eliminated by the test taker simply on the basis of general understanding. For example:

- Japan has few natural resources. To prosper and survive, the country must import raw materials, maintain high standards of manufacturing, and sell finished goods in foreign markets.

Japanese prosperity depends *most* on:

  a. discovering new raw materials
  b. importing manufactured goods
  c. her people's religion
  d. high levels of international trade*

In this item, distractor c. could be eliminated easily because it has no relationship to the text.

2. When the stems are incomplete statements that call for completion, use distractors that are grammatically compatible with the stem. For example, in the item on Japanese prosperity, if alternative (d) had read *to sell finished goods internationally,* the item would have been much more difficult and confusing to the test taker because the grammatical form of the alternative does not fit with the stem.

3. In principle, all distractors should be equally attractive and plausible to the test taker; that is to say, each distractor will be chosen equally often by test takers who do not choose the right answer. Distractors that are never or seldom chosen instead of the right answer are not serving any useful function. In practice, it is difficult to create distractors that are equally attractive, but some effort should be put into achieving this.

   When devising distractors, it might be helpful to (a) define the grammatical or semantic category to which the distractors should belong, and (b) think of alternatives that have some association with the stem or correct choice. In order to determine the attractiveness of distractors, it is necessary to keep a record of how often each distractor for a given question is chosen.

One method of identifying distractors is to choose them from the errors that students make in their spoken or written use of the language. Choosing distractors in this way means that you are likely to include plausible distractors that are attractive to the students who do not know the right answer.

4. Choose distractors that can tell you something about where the students are going wrong if they select them. A related point, avoid trick alternatives that distract the test taker for trivial or unimportant reasons. For example:

   - Definition item: to cook by exposing to direct heat

     a. roost
     b. burn
     c. broil*
     d. fry

   Roost is a trivial and tricky distractor because it confuses word meaning and pronunciation in a way that is not useful. It was chosen because of its resemblance to *roast.*

5. Choose distractors that have comparable difficulty, complexity, and length. Distractors that are obviously different from the alternatives might be especially salient to the test taker with the result that they are more likely to be eliminated or accepted. For example:

   Choose the best definition of the underlined word:

   - Mary is a very *bright* student; she got As in all of her courses.

     a. difficult
     b. erudite*
     c. shiny
     d. friendly

   Erudite is a poor alternative since it is a much more sophisticated word than the others and might be chosen for that reason alone.

6. Avoid including more than one correct alternative. In this regard, avoid distractors that might be correct in another dialect, regional variation, or modality of the language. For example, if you are testing spoken language, avoid using language in the distractors that might be considered appropriate in the written form of the language. An exception to this is if you want to test sociolinguistic skills. The best way to avoid more than one correct alternative is to have someone else review your test items.

   There should be no “missing link” between the stem and the alternative responses that would make more than one of them correct. This can happen when students assume some additional plausible context. For example:

- He left the office early ______ he could do some shopping.

  a. so*
  b. if*
  c. unless
  d. that

In this item, both *so* and *if* could be correct depending on the context you have in mind.

7. Avoid using alternative answers that contain words or phrases that match the stem if the other alternatives do not contain similar matching elements. An alternative that matches the stem while others do not might be chosen on the basis of the matching elements alone. In some cases this might lead to a correct choice, whereas in other cases it can lead to an incorrect choice. This is particularly important in comprehension tests.

8. It should not be possible to choose the correct response on the basis of general knowledge. In other words, choosing the correct response should depend on the content of the test. One way of examining this possibility is to have someone answer the questions without reading or hearing the text. For example:

- When tourists from Canada go to Florida on vacation, they travel ______.

  a. north
  b. west
  c. east
  d. south*

9. The alternative responses should be as simple as possible in keeping with the complexity of the test purposes. Avoid repetitious wording, redundancy, and unnecessary detail in the responses. For example:

- Robert went to the hospital

  a. because he wanted to visit his sick brother
  b. because he wanted to have his leg examined
  c. because he was a volunteer worker in the gift shop
  d. because his brother has asked him to

In this case, it would be better to include all of the repetitious elements in the stem: Robert went to the hospital because. . . .

10. In reading comprehension tasks, when several items are based on a single text, the answer to one question should not be given by the wording of another. For example:

    1. What did Mary serve Sam?

        a. leftover casserole
        b. scrambled eggs
        c. hamburger and fries
        d. fresh salmon*

    2. Where did Sam go for dinner?

        a. to Mary's*
        b. home
        c. to the school cafeteria
        d. to Joe's Restaurant

11. Avoid answers that refer to several other answer choices. For example:

    - She didn't go to the party because she _______

        a. was sick.
        b. had nothing to wear.
        c. was expecting an important call.
        d. b and c but not a.

Task 3

Have each student in the class write a multiple-choice question to test knowledge of some point in this chapter. Present your question to other students for review and feedback.

Assembling multiple-choice questions

The following are a number of points to take into consideration when putting multiple-choice items together for a test or examination.

1. Make sure the stem is distinct from the alternative answers. In written tests this can be achieved by inserting extra spaces between the stem and the alternative responses and by listing and indenting the alternatives on separate lines, as in the examples used in this chapter. In oral tests, the stem can be distinguished from the alternative responses by presenting the stem in one voice, say a female voice, and the alternatives in another voice, a male voice.

Poor presentation

- The population of Denmark is: (a) 2 million, (b) 4 million, (c) 7 million, (d) 15 million

Good presentation

- The population of Denmark is:

  a. 2 million
  b. 4 million
  c. 7 million
  d. 15 million

2. Identify the stems and alternatives using different symbols: for example, numbers for the stems and letters for the alternatives. When using separate answer sheets, make sure that your method of identifying stems and alternatives on the test corresponds to that presented on the students' answer sheet.
3. The correct alternative should occur equally frequently in each option position. Avoid presenting the correct choice in a particular position.
4. Use as many alternatives as are both possible and reasonable. The chances of selecting the correct alternative by guessing alone diminishes with more alternatives. With three alternatives, students have a 33% chance of getting the correct answer by guessing; with five alternatives, the chances of a correct response due to guessing is reduced to 20 percent. However, increasing the number of alternatives makes it increasingly difficult to construct plausible, attractive, and appropriate distractors.
5. Allow plenty of space between questions so that the test does not appear to be compressed and jammed together.

Summary

In this chapter, we discussed some factors to consider when choosing tasks to use in devising tests. We also presented guidelines for preparing closed-ended and open-ended test tasks. In the case of open-ended tasks, they took the form of general and specific questions to be asked about the demands posed by different tasks. Open-ended tasks can be easy to devise but time consuming to score. They often have the advantage of reflecting the way authentic language is used. In the case of closed-ended tasks, the guidelines we presented took the form of specific technical suggestions. It is important to recognize that the guidelines we presented for open-ended test tasks also apply to the preparation of closed-ended test tasks. Good closed-ended test tasks require considerable time and thought to prepare. Whether it is worth investing the time and thought needed to devise these kinds of tests depends on how the test results will be used and the importance of the decisions based on those results. Clearly, the investment of a great deal of time and thought is warranted when there are a large number of students to be tested. Another consideration when deciding whether to use a closed-ended test format is authenticity: arguably, many closed-ended test formats often do not reflect the way authentic language is used. They may nevertheless be useful for evaluating specific aspects of language learning. In some cases, closed-ended test tasks do call on the kinds of language performance your students will be expected to demonstrate. As in other aspects of classroom-based evaluation, one form of testing is not necessarily desirable under all circumstances and for all purposes. Rather, judicious use of each form may be called for.

Discussion questions

1. List as many possible language skills you can think of that can be assessed adequately using closed-ended response tasks. Now do the same for open-ended test tasks. In each case, limit yourself to five minutes. Then compare the kinds of language skills you have included in each category.
2. What are the advantages and disadvantages of using closed-ended tasks? open-ended tasks?
3. Devise a set of open-ended items for the same purpose. In order to do this, you will need to devise learning objectives associated with the content of this chapter. Try doing this individually, and then compare the objectives prepared by different students.
4. Select a multiple-choice test that you or others have used (this could be a standardized test), and then carefully examine the following aspects of the test using the guidelines suggested in this chapter: (a) the stems, (b) the alternative responses, (d) the instructions and answer sheet, and (e) the layout.
5. Are there other suggestions you would make for devising open-ended tasks in addition to those suggested in this chapter?
6. Have you ever devised a multiple-choice test? For what purpose? What did you find difficult about making it? What did you find useful about it? What were its limitations, if any, with respect to informing you about student achievement?

7. Select (or imagine) an open-ended test task you have used recently. How did you decide on the content and format of the test? On what basis did you devise your scoring scheme?

Readings

Airasian, P. (1991). Performance assessment. In P. Airasian (Ed.), *Classroom assessment* (pp. 251–306). New York: McGraw-Hill.

Annual Review of Applied Linguistics. (1995). *15.*

Carlson, S. B. (1985). *Creative classroom testing.* Princeton, N.J:. Educational Testing Service.

Carroll, B. J. (1980). *Testing communicative performance.* Oxford: Pergamon.

Carroll, B. J., and P. J. Hall. (1985). *Making your own language tests.* Oxford: Pergamon.

deJong, J. H. A. L., and D. K. Stevenson. (1990). *Individualizing the assessment of language abilities.* Clevedon, England: Multilingual Matters.

Fradd, S. H., P. L. McGee, and D. K. Wilen. (1994). *Instructional assessment: An integrative approach to evaluating students.* Reading, Mass.: Addison-Wesley.

Hamp-Lyons, E. (1991). *Assessing second language writing in academic contexts.* Norwood, N.J.: Ablex.

Hauptmann, P. C., R. Leblanc, and M. B. Wesche. (Eds). (1985). *Second language performance testing.* Ottawa: University of Ottawa Press.

Heaton, J. B. (1975). *Writing English language tests.* London: Longman.

Hughes, A. (1989). *Testing for language teachers.* Cambridge: Cambridge University Press.

Jacobs, H. L., S. A. Zinkgraf, D. R. Wormuth, V. F. Hartfiel, and J. B. Hughey. (1981). *Testing ESL composition: A practical approach.* Rowley, Mass.: Newbury House.

Madsen, H. S. (1983). *Techniques in testing.* New York: Oxford University Press.

Omaggio, A. C. (1983). *Proficiency-oriented classroom testing.* Washington, D.C.: Center for Applied Linguistics.

Underhill, N. (1987). *Testing spoken language: A handbook for oral testing techniques.* Cambridge: Cambridge University Press.

Valette, R. M. (1977). *Modern language testing.* New York: Harcourt Brace Jovanovich.

Weir, C. (1993). *Understanding and developing language tests.* New York: Prentice-Hall.