

2 TASKS—THE BUILDING BLOCKS OF TESTS

Introduction

This chapter should probably be viewed more as a list of some things to keep in mind when creating tests than as a list of options from which to randomly choose. It is intended to serve as a handy reference and source of ideas for different test tasks that teachers can use when developing tests, but this is not the final chapter in the book. Consequently, if you stop reading at the end of this chapter, your tests will look as though they had been written by someone who only read the first two chapters of a testing book. Chapters 3 and 4 will address how to plan the test itself; in particular, Chapter 3 details how to select appropriate tasks on the basis of what we are trying to assess and our reasons for giving the test. After all, you will not really know what task types you need until you have considered what your needs *are*, which is a crucial early part of the planning process. Nearly all of the test tasks you use on your own tests *will* look like the ones described here, but that does not mean that something is necessarily a bad idea merely because it is *not* included here. Similarly, just because something *is* described here does not mean it would necessarily be a good choice in your particular case.

Why Focus on Task Types at This Point?

The question naturally follows, then: Why include this section describing various task types now, rather than later, following the discussion of planning and choosing task types? In fact, I almost didn't! In the end, though, just as it is handy to know what sorts of building materials are available before creating the blueprints for a house, it seems useful to have teachers keep in mind the possibilities described here while planning their tests. This, therefore, is meant as a partial listing of the options that are available: the building blocks from which you can choose when you plan your tests. It is intended to help make new testers aware of what some of the possibilities are, and to indirectly suggest that "good" and "professional" do not always mean "multiple-choice."

Unfortunately, this last point actually needs to be emphasized. Many people, both teachers and students, are convinced that "real" tests of the serious, professionally developed sort must always be multiple-choice. Years ago, when I was teaching in

an intensive English program, I had to grade final exams and turn in course grades, in under 16 hours—preferably within *three* hours! I therefore created 100-question multiple-choice final exams for two of my courses. As little as I knew about testing at the time, I knew enough to be dismayed when someone who had never *seen* the tests described them as “wonderful.” After all, they were multiple-choice, and we had optically scanned answer sheets ... What more could anyone ask for? There are good reasons why standardized tests are almost always multiple-choice—although this may change in the near future (see Chapter 10)—but it does not always follow that all multiple-choice tests are good, nor that all good tests are multiple-choice.

On the other hand, automatically choosing a particular format simply because it is *not* multiple-choice would also be ill-advised. Similarly, looking through the tasks listed here and choosing one or two before doing any other planning would also be very unwise. To continue the house-building metaphor, it would be as bad an idea as buying some building materials and then starting to build a house, without having made any blueprints first.

Task Format

Before proceeding any further, it is important to establish the terminology that will be used here. By **task format**, I mean what a test task “looks like,” in terms of what sort of input test takers must process, and how we expect them to respond (for example, choosing an option, writing one word, or performing a role-play). Other writers have sometimes referred to this as “task type” (Bachman and Palmer 1996), “item format” (Brown and Hudson 2002), “test method” (Bachman 1990), “method” (Allen and Yen 1979), “response format” (Famularo 2008; Shin 2008), or “technique” (Alderson 2000). I have chosen “task format” for three reasons. First, it implicitly suggests that tests are not necessarily composed entirely of items, or questions. Second, it makes clear that we are considering the “shape” of a task, as opposed to what ability it is intended to assess. Finally, it emphasizes the point that the format includes *what* students are responding to, not just the *way* that they are supposed to respond.

When discussing task formats, there are three important types of classifications that we can make; these give a fairly good idea of what a given task's format will be. The first of these involves the type, or format, of response that a particular task calls for. **Selected response** tasks require the examinee to choose the correct answer from among several options, where “several” can mean as few as two, or as many options as can be squeezed onto one page. (Note that I am not recommending this “maximum squeeze” approach! I am only pointing out that this would be an example—a bad one—of selected response.) **Constructed response** tasks, on the other hand, call for the examinee to write or say something in response. Constructed response tasks can be further subdivided into **limited production** and **extended production** tasks (Bachman and Palmer 1996). Limited production tasks call for examinees to provide a short answer, which may be as short as one word, or as long as a sentence (if written) or an utterance (if spoken). Extended

production tasks require responses that are longer than one sentence or utterance. They may only be two sentences or utterances long, or they might be much more extensive.

The second way of categorizing task formats is the distinction between *items* and *prompts*. An **item** is a question that requires a short answer, or the selection of an appropriate response; thus, items are always selected response or limited production tasks. When we talk about “test questions,” items are what we usually have in mind. A **prompt**, on the other hand, usually asks the test taker to provide an extended response. This is the sort of task format most commonly used with speaking and writing assessments.

The third classification has to do with whether a test task is *passage-based* or *independent*. **Passage-based tasks** require examinees to read or listen to material in order to perform the task, as with reading comprehension questions. An important point with these tasks is that the questions cannot be answered, or tasks cannot be performed, without comprehending the passage. If test takers can do the tasks without reading or listening to the passage, after all, how can they be passage-based? This topic of ensuring **passage dependence**—that is, the need to comprehend a reading or listening passage in order to answer questions about it—is addressed further in Chapter 5, but suffice it to say for now that this fundamental point is neglected all too often by teachers writing tests of reading and listening comprehension. **Independent tasks**, on the other hand, do not require the processing of any additional material; test takers can respond to the task based solely on the item or prompt itself. Examples of independent tasks would be a grammar item that does not have any relationship to a reading or listening passage, or a speaking or writing prompt that is not based on a reading or listening passage.

Where Test Tasks Come From

Norris and his colleagues (Norris et al. 1998: 25) discuss two approaches to developing tests: *construct-based* and *task-centered*. (They actually state that these are two approaches to developing performance assessments, but the two approaches apply equally well to test development in general.) These approaches are somewhat reminiscent of McNamara’s strong and weak senses of performance assessment, but the two sets of distinctions are not the same, and are not necessarily even parallel.

Construct-based test development starts out by considering the construct that is being assessed, and selecting or creating test tasks that will tap into it; that is, tasks that will elicit evidence showing how much of the construct each examinee has. **Task-centered test development**, on the other hand, starts with deciding which tasks from the TLU domain—typically real-world language use tasks—are of greatest interest, sampling from among those tasks, and adapting those selected tasks for use on the test. Sampling is necessary because not everything can be included in the test. Similarly, adaptation is needed because we are usually simulating real-life tasks in a classroom, not following students around in the real world and scoring how they do “out there.”

In either case, as Bachman (2002b) points out, we need to keep in mind both constructs and the tasks which are used to assess them. Constructs, after all, are the whole point of the test. Test users make a decision—placement, achievement, etc.—based on what they believe about how much of the construct each student has. These inferences about constructs are based on test scores, which depend on how well students perform the tasks on the test—for example, how many questions they get right. At the same time, however, we should make our tasks authentic—that is, they should resemble TLU domain tasks as much as possible, at least to the extent permitted by available resources. Ignoring authenticity is highly problematic, because if test tasks do not resemble TLU domain tasks, this limits the degree to which we can generalize to that domain (Bachman and Palmer 1996). Thus, if we want to make statements about what examinees can do with language in the real world, the tasks they perform on the test cannot be entirely divorced from real-world tasks.

Focusing only on the task aspect leads to problems with how well we can extrapolate from test performance and make claims about performance on other tasks not actually included in the assessment. Ignoring the resemblance between test tasks and real-world tasks, and focusing on constructs alone, however, does not work either. This is why Bachman (2002b) argues that we must adopt both task- and construct-based approaches in our test design; that is, while we may start planning our tests from one or the other perspective, we must address both sets of issues.

Regardless of their origins, however, test tasks tend to fall into several predictable formats. Ways of classifying these formats were described above; what follows next is a discussion of what several of the more common formats are, what they are typically used to assess, and a few points to keep in mind when assessing particular constructs. Chapters 3 and 4 will then discuss in further detail the planning of which task formats to include in a given test, which depends on the specific ways in which we define the construct(s) we wish to assess on a given test, as well as on the TLU domain to which we wish to generalize.

Commonly Used Task Formats

Before going on to describe a number of task formats and their more common uses, it is important to point out that a particular task format should not be chosen merely because it looks interesting, nor merely because it seems easy to implement. This is not meant as an argument that teachers should plan tests that are boring and impractical, but rather that someone developing a test should attempt to make principled decisions. Novelty, convenience, and expediency are important, but they should not be our only concerns. As noted earlier, guidelines for selecting particular task formats are covered in Chapter 3. Chapter 4 then explains, once a particular task format has been chosen, how to write the specifications for it—that is, how many options there should be in a multiple-choice question, how long responses to a writing prompt should be, and so on. Advice on actually writing

the items and prompts, and creating or selecting reading and listening passages, is provided in Chapter 5. Finally, a task format should not be chosen simply because it is listed below; while I would certainly not argue that any of these task formats are inherently inappropriate options, that is not quite the same thing as arguing that a particular one is a good idea in a particular situation. Choose the tool that will best serve the purpose you have in mind.

Selected Response Tasks

Selected response tasks include some of the best-known task formats used in assessment. They are all item-based tasks, and can be used in association with reading or listening passages, or by themselves. Perhaps the best recognized of these task formats is **multiple-choice**. In multiple-choice tasks, as with any item-based task, examinees are presented with the **item stem**, the portion of the item that poses the question being asked. Below this are several **options**, *answer choices*, or *alternatives*, the possible answers from which the test taker must choose. The correct option is called the **key**, and the incorrect answers are the **distractors**. Multiple-choice items are popular for a number of reasons, including ease of scoring. Unfortunately, while they are easy to write, it often proves difficult to write *good* questions without training and experience. On the other hand, it is entirely too easy to write *poor* ones. (How to write better questions, how to identify stronger and weaker ones, and tips for improving poor items will be discussed in Chapter 5.) There are also concerns over how well this task format can assess language ability, although research also exists supporting its use. For example, in the area of reading comprehension, some research (for example, Rupp, Ferne, and Choi 2006) indicates that test takers read differently when taking multiple-choice tests than they would otherwise. On the other hand, Freedle and Kostin (1994) report three related studies which indicate that tests using this format can be construct-valid.

53 There ____ four chairs in the kitchen.

A has

C is

B have

D are

54 Eggs come ____ chickens.

A from

C inside

B with

D for

55 I like ____ game.

A it

C which

B this

Figure 2.1 Examples of Multiple-Choice Questions From a Test For Beginning-Level Students

- | | |
|----|---|
| 8 | Where is the conversation taking place? |
| A | <u>at a bookstore</u> |
| B | in a library |
| C | in a museum |
| D | at a university |
| 24 | The operator works for a _____ company. |
| A | telephone |
| B | game |
| C | <u>computer</u> |
| D | music |

Figure 2.2 Examples of Multiple-Choice Questions From an Intermediate-Level Listening Comprehension Test

True-false questions are essentially multiple-choice items with only two answers; aside from being easier than ordinary multiple-choice items to guess correctly, however, they are also considerably harder to write well, and often wind up becoming trick questions—that is, they tend not to have one unambiguously correct answer. One way that teachers sometimes try to counter this is by having students correct false statements to make them true, or underline the part of the stem that makes the item false; this, however, requires the items to be scored as limited production questions, which is a different task format. It also requires that the scoring key take into account how to treat questions that were true, but which are altered by students to make *other* true statements.

Another commonly used selected response task format is **matching**, a selected response format in which the correct option is accompanied by a relatively large number of distractors, many of which are the correct answers to other items. Matching can be a useful task format, but it is easy to go wrong with these tasks and make them more a matter of logic and the process of elimination, rather than tools for assessing the construct of interest. One way around this is to include more alternatives than there are questions or blanks. It should be pointed out that matching tasks are not limited to cases where students write the letter of the correct choice or draw a line connecting a list of words to a list of stems or pictures. They also include tasks where students write in a word from a list or word bank, which would therefore include information transfer tasks where the information to be written down is in a list. Matching tasks can be easy to write, but as with the other selected response formats, there are several ways in which it can be rather easy to go wrong. These will be discussed later in Chapter 5.

The final selected response format that I will discuss is **ordering tasks**, which ask examinees to arrange words, sentences, or paragraphs into the appropriate order. Such tasks may be useful for assessing sensitivity to rhetorical organization, for example, but they suffer from the same problem as many matching tasks—that is, unless there are more options than questions, it becomes easier to guess the correct answers as more and more questions are answered.

Selected response tasks probably owe most of their popularity to the ease with which they are scored. That is separate, of course, from the important question of

how *selecting* an answer compares with *providing* an answer. Research (for example, Brown 1980; Chapelle and Abraham 1990; Shohamy 1984; Wolf 1993) confirms what most of us might expect: selected response tasks tend to be easier than their limited production equivalents. The extent to which selected response might assess something qualitatively different from limited production, however, is less clear. Selected response tasks will also tend to be less authentic than their limited production equivalents.

Limited Production Tasks

Like selected response tasks, limited production tasks are typically item-based, and can be either passage-based (for example, reading or listening comprehension items) or independent (for example, various types of short-answer grammar questions). Most commonly, these are short-answer questions, although other task formats are possible, such as information transfer tasks *without* a list of the words or phrases that students are supposed to write. Other examples of task formats besides short-answer questions include fill-in-the-blank (also known as *gap-fill*) items, usually one sentence long; sentence writing or combining tasks, such as when students are given a set of words with which to form a sentence; completing a blank or incomplete graphic organizer; and completing an incomplete outline. These two tasks may be used to assess sensitivity to the rhetorical structure of a passage (for example, Vongpumivitch 2004), or the ability to recognize main ideas, major ideas, and supporting details (for example, Shin 2008). Limited production tasks require a written response, ranging in length from one word to a phrase, or perhaps as much as a sentence. It is also possible to have spoken or written questions that require a short spoken response, but this is much less common. Good limited production items are generally somewhat easier to write than good selected response items, but they pose two difficulties of their own, both involving scoring.

Short-answer item:

- 1 What did Alice follow down the hole?
a rabbit

Fill-in-the-blank item (revised version of short-answer item above):

- 1 Alice followed a(n) rabbit down the hole.

Sentence combining item:

- 17 *Lions / not / eat / vegetables*
Lions do not eat vegetables.

Figure 2.3 Examples of Some Limited Production Tasks From a Reading Comprehension Test

Classifying Polygons	Classifying Polygons
I Triangles have three sides A _____ B Isosceles triangles have two equal sides C _____	I Triangles have three sides A <u>Right-angle triangles have one 90° angle</u> B Isosceles triangles have two equal sides C <u>Equilateral triangles have three equal sides</u>
II _____ A Irregular quadrilaterals have no parallel sides B Trapezoids have two parallel sides C _____ 1 _____ 2 Rhombus: a parallelogram with four equal sides 3 Rectangle: a rhomboid with four right angles 4 Square: a rhombus with four right angles	II <u>Quadrilaterals have four sides</u> A Irregular quadrilaterals have no parallel sides B Trapezoids have two parallel sides C <u>Parallelograms have two pairs of parallel sides</u> 1 <u>Rhomboid: a parallelogram with unequal adjacent sides and no right angles</u> 2 Rhombus: a parallelogram with four equal sides 3 Rectangle: a rhomboid with four right angles 4 Square: a rhombus with four right angles
III Pentagons have five sides	III Pentagons have five sides
IV _____	IV <u>Hexagons have six sides</u>
V _____	V <u>Heptagons have seven sides</u>
VI Octagons have eight sides	VI Octagons have eight sides

Figure 2.4 Example of an Incomplete Outline Task From a Reading Comprehension Test

The first issue, and probably the most widely recognized one, is that short-answer questions take longer to grade or score than selected response items, making them less practical for testing beyond the level of the individual classroom, as well as in cases where a teacher has a large number of students, little time in which to grade tests, or both. This is especially true when the desired answers are longer than a single word or phrase, although this is becoming less of a problem with the advent of **web-based testing (WBT)** and increasingly practical methods for **computer automated scoring (CAS)**, as will be discussed in Chapter 10.

The second problem has to do with the scoring key. As a rule, most questions will have more than one possible correct answer, particularly when learners are above the most basic level. Therefore, the scoring key needs to include all synonyms or other alternative answers that are considered acceptable. If only one teacher is doing all the scoring, this becomes less necessary; on the other hand, the larger the number of tests being scored, the more important it is to write these decisions down, for the sake of consistency. This is even more true when partial credit (i.e. less than full credit, or fewer points than the maximum possible) is being awarded for certain answers that are deemed partially correct, but not adequate for receiving full credit. It might seem that experienced teachers should be able to apply common sense regarding which responses are close enough and which ones are not. Unfortunately, this does not work as well as one might expect: Teachers often disagree, sometimes strongly! Since it should not matter to a student which teacher happens to be grading their paper, luck should be removed from the equation through a carefully specified scoring key.

Deletion-Based Tasks

Deletion-based tasks involve taking a written passage and deleting words or portions of words, which are then replaced with blank spaces. Test takers must then fill in these blanks, either by writing in the empty space or by selecting from a range of options. They are most commonly a subset of limited production items, but they can also be written using selected response formats. For this reason, I will treat them as a separate category of task format.

Cloze Tests

Probably the best-known, most prototypical type of deletion-based task is the **cloze test**. In a cloze test, single words throughout a passage are replaced with blanks, and students must write the missing words in the blanks, or select the correct word in the case of multiple-choice cloze. Normally, the first two sentences or two will be left intact, so readers can gain a sense of the passage. In **fixed-deletion cloze**, the words are deleted randomly—that is, every n th word is deleted, where n is some number chosen in advance. For this reason, this task format is also known as *n th-word deletion cloze* or *random deletion cloze*, since the deletions are at evenly-spaced intervals, which makes it effectively random which words (and which *types* of words) will be deleted. Obviously, the fewer words between deletions, the harder

it will be to answer correctly; therefore, n is usually at least 5 (Alderson 1983), but often 10 or more.

Fixed-deletion cloze in second language testing was originally proposed as a means of assessing overall language proficiency (Oller 1979). Subsequent studies have found that it can assess grammatical competence in some cases, and textual competence—that is, understanding of cohesive ties—in others (Chapelle and Abraham 1990). It may be that these inconsistent results are because of the random nature of the deletions; if so, Bachman (1985) points out that switching to rational deletion (see below) would allow teachers to assess the specific areas in which they are interested.

Original Passage (Carroll 1865)

Alice opened the door and found that it led into a small passage, not much larger than a rat-hole: she knelt down and looked along the passage into the loveliest garden you ever saw. How she longed to get out of that dark hall, and wander about among those beds of bright flowers and those cool fountains, but she could not even get her head through the doorway, “and even if my head *would* go through,” thought poor Alice, “it would be of very little use without my shoulders. Oh, how I wish I could shut up like a telescope! I think I could, if I only know how to begin.” For, you see, so many out-of-the-way things had happened lately, that Alice had begun to think that very few things indeed were really impossible.

Random Deletion Cloze ($n = 7$)

Alice opened the door and found that it led into a small passage, not much larger than a rat-hole: she knelt down and looked along the passage into the loveliest garden you ever saw. How she longed _____ get out of that dark hall, _____ wander about among those beds of _____ flowers and those cool fountains, but _____ could not even get her head _____ the doorway, “and even if my _____ *would* go through,” thought poor Alice, “_____ would be of very little use _____ my shoulders. Oh, how I wish _____ could shut up like a telescope! _____ think I could, if I only _____ how to begin.” For, you see, _____ many out-of-the-way things had happened lately, _____ Alice had begun to think that _____ few things indeed were really impossible.

Rational Deletion Cloze (prepositions, conjunctions, relative pronouns, and other connectors)

Figure 2.5 Examples of Some Deletion-Based Tasks (note that these examples are briefer than would be a good idea in real life)

Alice opened the door and found that it led into a small passage, not much larger than a rat-hole: she knelt down and looked along the passage into the loveliest garden you ever saw. _____ she longed to get out _____ that dark hall, and wander about _____ those beds of bright flowers _____ those cool fountains, _____ she could not even get her head _____ the doorway, "and even _____ my head *would* go through," thought poor Alice, "it would be _____ very little use _____ my shoulders. Oh, how I wish I could shut up _____ a telescope! I think I could, _____ I only know _____ to begin." For, you see, so many out-of-the-way things had happened lately, _____ Alice had begun to think _____ very few things indeed were really impossible.

C-Test (n = 7; "odd" letters deleted; blanks shifted one to the left for one-letter words)

Alice opened the door and found that it led into a small passage, not much larger than a rat-hole: she knelt down and looked along the passage into the loveliest garden you ever saw. How she longed t_____ get out of that dark hall, a_____ wander about among those beds of bri_____ flowers and those cool fountains, but s_____ could not even get her head thr_____ the doorway, "and even if my he_____ *would* go through," thought poor Alice, "i_____ would be of very little use wit_____ my shoulders. Oh, how I wi_____ I could shut up like a tele_____! I think I could, if I only kn_____ how to begin." For, you see, s_____ many out-of-the-way things had happened lately, th_____ Alice had begun to think that ye_____ few things indeed were really impossible.

Figure 2.5 Examples of Some Deletion-Based Tasks (note that these examples are briefer than would be a good idea in real life) (continued)

Aside from questions about what, exactly, it measures, another issue with fixed-deletion cloze involves the starting point for deletions. Specifically, changing the exact position at which the deletions start may lead to large changes in the scores on two cloze tests based on the same passage, even with as many as 50 blanks. Porter (1978) created two tests based on the same passages, with the only difference between them being that deletions in one test occurred one word earlier than in the other. Scores on the two tests had surprisingly low correlations, indicating that to a large extent, the two were measuring substantially different things. Subsequent research by Sciarone and Schoorl (1989) found that if this problem is to be avoided, about 75 blanks is the minimum number of blanks if acceptable-word scoring is used, and 100 blanks for exact-word scoring (see below for an explanation of these two types of scoring). This presumably has to do with adequacy of sample size: Since words are being randomly selected for deletion, a certain number have to be chosen for a given "sample" to be representative of the text as a whole. Two *unrepresentative* samples of words, after all, can hardly be expected to be comparable. Thus, for fixed-deletion cloze, if we want the minimum of 75 blanks needed to ensure representative sampling of various language features (assuming n is from five to ten words) a passage of at least 375–1,000 words would be required.

A second approach to cloze testing is **rational-deletion cloze**, also referred to sometimes as *gap-fill* tasks, in which words of a specific type (for example, connectors, nouns, or prepositions) are deleted. Of course, deletions should not be clustered together too closely, as sufficient context needs to be provided for test takers to be able to infer what belongs in each blank. One way to avoid this is to set a target for the number of deletions in a passage, and then spread them more-or-less evenly throughout, leaving the first sentence or two of the passage intact. Much like fixed-deletion cloze, it is helpful to start from the end of the passage and work backwards, but trying to delete the target words at *roughly* similar intervals; obviously, *identical* intervals, as used in *n*th-word deletion cloze, will not be possible here.

Rational-deletion cloze can be used to assess a number of different constructs, including vocabulary (for example, deleting nouns), cohesion (for example, deleting various cohesion markers, including content words in some cases), or various aspects of grammar (for example, deleting prepositions). Indeed, many common grammar exercises are arguably modified versions of rational-deletion cloze, such as passages in which blanks are inserted before every noun, with *ø* to be inserted when no article is required. Another example is passages with infinitives in parentheses next to each deleted verb; students must then conjugate the verbs based on the context provided by the passage. Obviously, these tasks can be used for assessment as well.

By definition, the blanks in rational-deletion cloze are not a random sample of the words in the passage. Thus, two rational-deletion cloze tests based on the same passage and targeting the same construct(s) will probably have similar patterns of deletions. Scores on the two tests will also probably be very similar, since they are assessing the same things. This presents an advantage over fixed-deletion cloze, in that it is not necessary to have such a large number of deletions.

A variant of rational-deletion cloze, which I will label **summary cloze**, is intended to assess reading comprehension. It involves writing a summary of a reading passage, and choosing the deletions that would be required to demonstrate comprehension of the original passage; that is, each word deleted provides information that could serve as the focus of a multiple-choice or open-ended comprehension question.

As noted above, cloze tasks can also employ selected response, either multiple-choice or matching. Deletions for selected response cloze tests can be fixed or rational, depending on the intended purpose of the assessment. Of course, this raises the question of whether limited production or selected response is more appropriate in a particular situation, a question that teachers should consider carefully before making a decision.

Aside from how the deletions were made, cloze tests are also distinguished by the method used to score them. **Exact word scoring** requires that test takers supply the word that was in the original passage; any other word, regardless of how well it might fit in the blank, is counted wrong. This arbitrary method poses serious issues of fairness, of course, since in many cases answering correctly might be a

question of luck, not of language ability. Language testing expert J. D. Brown perhaps phrases it best: “There is something inherently repugnant about counting an answer wrong, which is actually correct, simply because the author of the original passage did not choose to use that word” (Brown 1980: 316). In contrast, **acceptable response scoring** is an approach that counts any reasonable response as correct. It is also possible to expand this to include partial credit for incorrect answers that are close to correct; how close is “close enough” must be decided in advance by the test writer or test development committee.

Research by Brown (1980) comparing exact-word, acceptable-word, and multiple-choice scoring indicates that multiple-choice cloze tends to be the easiest of the three, but discriminates the most poorly between high- and low-ability test takers. Exact-word scoring is the most difficult of the three, and discriminates only slightly better than multiple-choice cloze. Acceptable-word scoring, however, does the best job of differentiating between high- and low-ability examinees, with an average difficulty level between those of the other two methods.

C-Tests

A **C-test** is similar to a cloze test, but with the second half of every n th word deleted. As with cloze tests, the very beginning of the passage should be left intact to give examinees a sense of the passage. C-tests are most commonly used as measures of general language proficiency—at least, in terms of written language—and in particular, seem to assess both vocabulary knowledge and grammar (Eckes and Grotjahn 2006). Some studies have further indicated that they assess knowledge of derivational and inflectional morphology (Chapelle 1994) and the strategic use of vocabulary knowledge and semantic associations (Babaii and Ansary 2001). In a cloze method comparison study, Chapelle and Abraham (1990) found that scores on a C-test were higher than on fixed- or rational-deletion cloze tests based on the same passage. C-test scores were also the most highly correlated with scores on a multiple-choice vocabulary test, as well as with a test of **field independence** (a cognitive style associated with more analytic processing, as opposed to **field dependence**, which is more holistic). (For a general discussion of field independence/dependence and its relationship to second language acquisition, see Chapelle and Green 1992; and Johnson, Prior, and Artuso 2000.)

Extended Production Tasks

Extended production tasks, as explained earlier, are intended to elicit responses that will be longer than a single sentence or utterance. They are almost always prompt-based, whereas limited production and selected response tasks are overwhelmingly item-based. Extended production tasks can be either *independent* or *passage-based*. An essay prompt, whether it is a short single-sentence question or an entire paragraph, is one very common example of an independent extended production task. Passage-based extended production tasks are, of course, integrated tasks by definition; some examples are provided in Table 2.1. Regardless of the specific task, it is important to specify the scoring criteria and procedures before

the test is administered. This will be discussed later in Chapter 7. I will divide extended production tasks into three categories here: those with *written* responses, those with *spoken* responses, and **portfolios**.

Nature of the integration	Example of a possible response
Listening and speaking	giving a short spoken reaction
Listening and writing	writing a summary of the listening passage
Reading and speaking	giving a spoken summary of the passage
Reading and writing	writing an essay reacting to the passage

Table 2.1 Examples of Passage-Based (i.e. Integrated) Extended Production Tasks

Extended Production Tasks With Written Responses

Written responses to extended production tasks can range in length from several sentences or a single paragraph up to an entire essay; in non-test assessments, where students have time to work at home, they can even include full-length research papers. These tasks include more than traditional writing prompts, however. One example of another task format is **recall tasks**, in which students write as detailed a summary as they can remember after listening to or reading a passage. In the case of reading, they are often allowed to read the passage as many times as they wish before putting it away and beginning their responses. In the case of a listening recall task, they might be permitted to listen more than once, or they might be allowed to take notes during the listening and then review them prior to writing. Recall tasks are more commonly used as research tools than for academic purposes, since they involve not only language ability, but memory as well. (In fact, many native speakers—including this author, in his native language—might even find such a task difficult!)

Written summaries of reading or listening passages are another extended production format, and are just what the name would indicate. Their primary difference from recall tasks is that test takers are allowed to refer to the passage (or their notes, in listening) when writing. Because they do not depend on a learner's memory, they are more suitable for curriculum-related assessment purposes. Not surprisingly, research by Riley and Lee (1996) indicates that students writing summaries provide more main ideas than students performing recall tasks. Interestingly, recall tasks seem to yield responses where a higher proportion of the ideas are details, as opposed to main ideas. For summary tasks, however, the pattern is reversed, with students' responses containing more main ideas than details. Given these patterns, then, any time we deem comprehension of main ideas to be more important than the ability to remember specific details—whether important ones or not—summary tasks are probably more appropriate.

Dictation is another familiar and transparent term, and integrates listening ability with writing ability; thus, if the ability to write down what a student hears is of interest, it may be a useful task format. Aside from its use as a pedagogical task

(see, for example, Natalicio 1979; Reinders 2009), dictation has also been used to assess knowledge of lexical “bundles” (Nekrasova 2009) and responsiveness to instruction on a particular syntactic feature (for example, Sheen 2009), and has been viewed by many (for example, Fouly and Cziko 1985; Oller 1972; Stansfield 1977, 1985; Young 1987) as an indicator of overall language proficiency. *Note-taking* is a related task format that may be appropriate in academically oriented language programs. Its main difference from dictation is that the goal of note-taking is not, of course, to write down every word that was heard. Two common questions faced when creating a dictation or note-taking task are how many times the passage should be repeated, and how quickly or slowly it should be read. The answers depend to a large extent on the generalizations about language ability that teachers wish to make. Thus, if examinees could listen to the passage multiple times outside the test, then it would be more authentic for them to hear it multiple times during the test as well. Likewise, if they would only hear it once in the real world, or in the classroom outside testing situations, then it would be more authentic to only let them listen once during the test. Authenticity must be balanced, though, with the practical matter of how many times most students will *need* to listen in order to be able to write what they heard. It is also important to consider the extent to which students will need extra time to write—even native speakers may have difficulty keeping up with fast speech.

Suggestions for developing scoring rubrics for recall, summary, dictation, and note-taking tasks will be discussed in Chapter 7.

Extended Production Tasks With Spoken Responses

Spoken responses to extended production tasks can vary greatly in length, depending on the amount of preparation time provided, the amount of input about which learners are to speak, and the format of the task. They may be as short as a few utterances (essentially, “spoken sentences”), a short monolog of 30 seconds, or a very short conversation. They can also include a lengthy conversation, or a speech or presentation of the same length as a “real” one in the TLU domain. Of course, the appropriate length will vary depending on the TLU domain task on which the assessment is based. When appropriate “real world” tasks are too long to be practical for use in classroom or other assessment settings, it may be appropriate to simulate a specific *portion* of the longer task.

The most prototypical extended production speaking task is perhaps the *interview*. Interviews can be highly scripted, with specific prompts which must be asked, or from which interviewers must choose. Follow-up questions may not be allowed, or may be required; in the latter case, the prompts serve more as jumping-off points for the interview. Other interviews are much more free-formed, however (for example, the ACTFL Oral Proficiency Interview; see Swender et al. 1999), with no set list of prompts from which to start. In more open-ended interviews, however, it is necessary to specify what sorts of things (topics, functions, structures, etc.) should be discussed or elicited. More open-ended formats also, naturally, require more training and experience to perform properly.

It has long been recognized (see, for example, van Lier 1989) that in many ways, interviews are not necessarily reflective of normal conversation. Most spoken interaction in the real world does not consist of Person A asking Person B a series of questions, with Person A selecting all the topics to be discussed, and making all the decisions about when to move on to the next topic. This does not necessarily mean that the interview is fatally flawed as a means of speaking assessment, but this issue of discourse authenticity has led to interest in alternative approaches to speaking assessment.

One such approach is the **group oral interview** (Folland and Robertson 1976), in which pairs or small groups of students discuss one or more specific topics; these may be a list of questions or topics, or a reading or listening passage. How small is "small enough" to be manageable is a question that will depend on a number of factors, such as the proficiency levels of the students, differences between their levels, their talkativeness, their shyness, how long the interview will last, how many raters are available to score the interactions, and how many interviewees a rater can evaluate simultaneously. Whether the interviewer(s) should be involved in the discussion—and if so to what extent—is another consideration to be borne in mind. Finally, although the group oral obviously offers the benefit of greater efficiency, and may permit the observation of discourse features not likely to emerge in traditional interviews, there are also concerns about the validity of score interpretations resulting from its use.

In summarizing research on the group oral, testing researcher Gary Ockey concludes (Ockey 2009b) that studies have presented mixed results. I interpret this as indicating that like most task formats, the group oral format *can* be used to produce useful results, but—like *any* format—does not possess any magical "inherent validity" that guarantees good results in every case. Ockey recommends that to make the group oral more useful, it must be implemented in a way that will be appropriate for the testing context, and students taking it need to be familiar with the testing format. He also notes that the complex nature of this test format means that a number of variables can influence examinee performance in ways that are still poorly understood, with sometimes unexpected results. (For example, Ockey's study finds that assertive test takers tend to score more highly when grouped with less-assertive test takers than when they are grouped with other assertive examinees. On the other hand, non-assertive test takers' scores are *not* affected by the assertiveness of other group members.) Thus, given the complexity and incomplete understanding of the variables involved in this task format, it should probably be used with caution in high-stakes settings, although that is not to say that it should never be used in such settings. In classroom settings, however, group testing may be the only way to make speaking assessment possible.

Another speaking task format closely related to the group oral is *role-play tasks*; in fact, it could be argued that it is merely an extension of the group oral interview. Role-plays provide the benefit of assessing learners' ability to engage in various types of speech acts, or to perform a wider variety of language functions, than would be the case in a discussion. Role-plays also make it possible to assess the

ability to speak in a wider range of situations, and about a wider range of topics. They are particularly appropriate if students are used to engaging in role-plays in class, as there are fewer concerns than with task format familiarity. Role-plays are usually easier to plan for two speakers than for larger numbers, although careful planning and a bit of creativity can help teachers increase the number of students to be included in a given interaction. Roles and background information must be clearly specified, of course, so that examinees will know how they are expected to perform. This is often done by providing the information on note cards, which the other students are not allowed to see. The language used on these role cards, of course, should be at a level equivalent to or below the language the examinees are expected to be producing, so that reading ability does not interfere with someone's ability to display their speaking ability.

When writing these role cards, of course, it is crucial that the directions be clear, specific, and understandable to the students. Nothing ruins a good role-play quite as much as having one of the students doing the opposite of what was intended. When a restaurant is supposed to be out of the special of the day, and the "waiter" was instructed to suggest another dish instead, it can ruin the entire task if the waiter instead simply takes the order as given. The "customer" then has no chance to ask questions about the rest of the menu, complain, demand to see the manager, or engage in any other speech acts beyond saying "Thank you."

Another somewhat simpler speaking task format is the *monolog*, which is often employed in speaking assessments in which responses are recorded (for example, the Internet-Based TOEFL, or iBT; Educational Testing Service 2009). Students may prepare a monolog in advance in some cases (for example, when they are required to memorize a poem beforehand), but memorized speeches or poems may be of limited value in terms of generalizing to communicative language ability. Thus, students are more frequently asked to respond to a prompt or prompts without having seen them in advance. These prompts may take the form of a number of different tasks and functions, including giving opinions, one-person role-plays (for example, "Imagine that I am your friend, and you want to ask me to come to your house to help you move some furniture. Call me on the telephone and ask me to help you."), narrating the story told by a series of pictures, describing a picture, or interpreting graphs. (Note that these tasks can be used in interview-based tests as well.) One important consideration for those planning monolog-based speaking tests is whether to allow time (ranging perhaps from a few seconds to one minute) for planning an unrehearsed response. The rationale for such time is that even in one's first language, it may take time to process pictures or graphs, consider opinions on an issue, and so forth.

Outside of certain Language for Specific Purposes (LSP) contexts, graph and picture comprehension is not something language teachers want to test, *per se*; rather, graphs and pictures are tools for eliciting samples of language. Xi (2005) recommends against including graphs with too many visual "chunks," particularly when test takers are not all used to reading graphs. She further recommends allowing planning time to help compensate for some examinees' lack of graph

familiarity. Naturally, this advice would apply equally to interview-based tests that included graphs, picture narration tasks, and so forth.

A final type of speaking task format that includes extended production responses is *reading aloud*, also referred to as oral reading. This task format is perhaps best suited to the assessment of pronunciation ability (see, for example, Celce-Murcia, Brinton, and Goodwin 1996)—including both segmentals (vowels and consonants) and suprasegmentals (stress, rhythm, intonation, connected speech, etc.)—and reading fluency (especially in first language reading; see, for example, Crosson and Lesaux 2009; Deeney 2010; Wanzek et al.). My personal experience in the classroom has also indicated that it can provide an indication that second language learners have difficulty with derivational and inflectional suffixes. On the other hand, it seems unlikely to provide much useful information about an examinee's command of grammar (aside from inflection), vocabulary, or any other components of communicative language ability.

Portfolios

Portfolios are collections of student work from a course, and are useful for showing a student's development and improvement. Portfolios are most commonly associated with writing, but they can be used in the assessment of speaking as well, as will be discussed below. Portfolios of student writing are probably already familiar to many teachers, and provide a more comprehensive picture of a student's writing ability than can be gotten from performance on a single timed occasion. Writing portfolios can show the development of a student's writing ability. They also provide a means of assessing students without having to worry about how test anxiety might impede performance. When pre-writing and drafts are included in the portfolio, they can provide a way to prevent—or at least detect—plagiarism. For example, a poor outline and three mediocre, barely-improving drafts of a paper, followed by a perfect final draft, may be cause for suspicion. A perfect final paper on an entirely different topic is even more suspicious.

On the other hand, writing portfolios are not without their drawbacks. For example, they do not reveal anything about how well a student can perform with limited time; this can be addressed, of course, by also using timed writing assignments to assess students. Perhaps the most serious issue, however, is that sometimes it can be difficult to tell how much of the portfolio is indicative of students' independent writing ability, as opposed to what they can do with assistance. This depends, of course, on the amount and nature of the assistance provided to students, and even when that assistance is considerable, portfolios can still be quite useful. For example, teachers can learn about student achievement by examining whether pieces from later in the course "evolved" into finished products more quickly than those written earlier. Similarly, comparing the quality of final drafts from early and late in the course can provide information on achievement (admittedly, scaffolded achievement), even if drafts are not included in the portfolio.

Although they seem to be more commonly associated with writing, portfolios can also be useful for assessing student progress and achievement in speaking ability.

The process is somewhat more involved than with writing portfolios, of course, since recordings of student speech must be made and organized. If students can record their speech digitally, this is much more feasible. In many cases, it may be necessary to train teachers and students in how to record, convert, copy, etc.

Whatever the recording format being used, it is important to specify to students what should be included, in terms of content and length. Options include recording of regular spoken journal “entries” (with or without assigned topics or prompts), interactions with native speakers or other students outside class, and even speaking activities in class. When recordings are being made of monologs, or student-to-student dialogs, it is important to keep in mind that these will generally be recordings of the students’ *best* work. Students should therefore be encouraged to re-record themselves if they catch errors, but not to repeat the process so many times that they begin to sound unnatural, as if they were delivering a recited performance.

Regardless of whether they are spoken or written, it is essential to keep in mind *why* a portfolio is being used for assessment—that is, what information is it expected to provide about a student’s language ability? As with any of the task formats described in this chapter, that question should be answered before the tasks are chosen.

Translation as a Task Format

Translation may fall under limited or extended production, depending on the length of the material to be translated. Probably the most important thing to keep in mind, though, is that translation is a good way to assess the ability to translate. As a means of assessing reading, writing, speaking, listening, grammar, vocabulary, discourse competence, or sociolinguistic competence, however, it is probably somewhat questionable. It can potentially *involve* any or all of these areas, but by virtue of that very fact, it is difficult to make defensible claims about any particular one of them.

Summary

In this chapter, we have looked at several ways to classify task formats. One of these is the distinction between selected and constructed response tasks, with constructed response further broken down into limited and extended production. We also considered the differences between items and prompts, and passage-based versus independent tasks. We compared task- and construct-based approaches to test design, noting that both TLU domain tasks and **construct definitions** must be kept in mind, regardless of which we approach we say we are following. Now that we have reviewed some of the options available when making a test, the next chapter will take up the question of how to choose from among them. This will be done as part of the test planning process, once decisions have been made about test purpose, constructs, TLU domains, and requirements for test usefulness. Chapter 4 will then address what specific details need to be specified about each test task—essentially, a plan for exactly how to write the test—and Chapter 5 will

provide advice for what to do while writing various types of items and prompts or selecting or creating listening and reading passages, as well as how to identify which questions in item-based tests need improving.

Further Reading

Bachman, L. F. 1990. *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

For a discussion of input, response, and the relationship between them, see Chapter 5.

Bachman, L. F. 2002b. 'Some reflections on task-based language performance assessment'. *Language Testing* 19: 453–476.

Bachman, L. F. and A. S. Palmer. 1996. *Language Testing in Practice*. Oxford: Oxford University Press.

For more on classifying input and responses, see their test task characteristics framework on Chapter 3, particularly the sections dealing with characteristics of the input and expected response.

Brown, J. D. and T. Hudson. 2002. *Criterion-Referenced Language Testing*. Cambridge, UK: Cambridge University Press.

See Chapter 3 for a discussion of task formats, which the authors refer to as item types.

Celce-Murcia, M., D. M. Brinton, and J. Goodwin. 1996. *Teaching Pronunciation: A Reference for Teachers of English to Speakers of Other Languages*. Cambridge, UK: Cambridge University Press.

Chapter 12 deals with testing and evaluation in the English pronunciation classroom, and Appendix 13 includes a one-page diagnostic passage for learners to read aloud.

Norris, J. M., J. D. Brown, T. Hudson, and J. Yoshioka. 1998. *Designing Second Language Performance Assessments* (Technical Report 18). Honolulu: University of Hawai'i, Second Language Teaching and Curriculum Center.

Discussion Questions

- 1 What are some examples of testing situations where (a) passage-based items and (b) independent items might be more appropriate?
- 2 What would you say are some of the benefits and drawbacks of using multiple-choice and short-answer items to assess the following areas of ability?
 - a Reading comprehension
 - b Listening comprehension
 - c Writing ability
 - d Speaking ability
 - e Grammatical competence
 - f Vocabulary knowledge

- 3 For a language class which you have taught, or have taken yourself, think of an achievement test (for example, a chapter or unit test, midterm exam, or final exam), and how part of it could be converted to a rational-deletion cloze test:
 - a What were some of the constructs (for example, grammatical structures, lexical categories, reading strategies) the test was supposed to assess?
 - b What category of words might you delete from a cloze passage in order to assess this construct?
- 4 In general, what do you think are some of the advantages and disadvantages of using each of the following response formats to assess grammar and vocabulary?
 - a Selected response
 - b Limited production
 - c Extended production