TESTING

LISTENING COMPREHENSION

CHAPTER 6

ASSESSING LISTENING

In earlier chapters, a number of foundational principles of language assessment were introduced. Concepts like practicality, reliability, validity, authenticity, washback, direct and indirect testing, and formative and summative assessment are by now part of your vocabulary. You have become acquainted with some tools for evaluating a "good" test, examined procedures for designing a classroom test, and explored the complex process of creating different kinds of test items. You have begun to absorb the intricate psychometric, educational, and political issues that intertwine in the world of standardized and standards-based testing.

Now our focus will shift away from the standardized testing juggernaut to the level at which you will usually work: the day-to-day classroom assessment of listening, speaking, reading, and writing. Since this is the level at which you will most frequently have the opportunity to apply principles of assessment, the next four chapters of this book will provide guidelines and hands-on practice in testing within a curriculum of English as a second or foreign language.

But first, two important caveats. The fact that the four language skills are discussed in four separate chapters should in no way predispose you to think that those skills are or should be assessed in isolation. Every TESOL professional (see *TBP,* Chapter 15) will tell you that the *integration of skills is of paramount importance in language learning.* Likewise, assessment is more authentic and provides more washback when skills are integrated. Nevertheless, the skills are treated independently here in order to identify principles, test types, tasks, and issues associated with each one.

Second, you may already have scanned through this book to look for a chapter on assessing grammar and vocabulary, or something in the way of a **focus on form** in assessment. The treatment of form-focused assessment is not relegated to a separate chapter here for a very distinct reason: there is no such thing as a test of grammar or vocabulary that does not invoke one or more of the separate skills of listening, speaking, reading, or writing! It's not uncommon to find little "grammar tests" and "vocabulary tests" in textbooks, and these may be perfectly useful instruments. But responses on these quizzes are usually written, with multiple-choice selection or fill-in-the-blank items. In this book, we treat the various linguistic forms

(phonology, morphology, lexicon, grammar, and discourse) within the context of skill areas. That way, we don't perpetuate the myth that grammar and vocabulary and other linguistic forms can somehow be disassociated from a mode of performance.

OBSERVING THE PERFORMANCE OF THE FOUR SKILLS

Before focusing on listening itself, think about the two interacting concepts of **performance** and **observation**. All language users perform the acts of listening, speaking, reading, and writing. They of course rely on their underlying competence in order to accomplish these performances. When you propose to assess someone's ability in one or a combination of the four skills, you assess that person's *competence,* but you observe the person's *performance.* Sometimes the performance does not indicate true competence: a bad night's rest, illness, an emotional distraction, test anxiety, a memory block, or other student-related reliability factors could affect performance, thereby providing an unreliable measure of actual competence.

So, one important principle for assessing a learner's competence is to consider the fallibility of the results of a single performance, such as that produced in a test. As with any attempt at measurement, it is your obligation as a teacher to **triangulate** your measurements: consider at least two (or more) performances and/or contexts before drawing a conclusion. That could take the form of one or more of the following designs:

- several tests that are combined to form an assessment
- a single test with multiple test tasks to account for learning styles and performance variables
- in-class and extra-class graded work
- alternative forms of assessment (e.g., journal, portfolio, conference, observation, self-assessment, peer-assessment).

Multiple measures will always give you a more reliable and valid assessment than a single measure.

A second principle is one that we teachers often forget. We must rely as much as possible on *observable* performance in our assessments of students. Observable means being able to see or hear the performance of the learner (the senses of touch, taste, and smell don't apply very often to language testing!). What, then, is observable among the four skills of listening, speaking, reading, and writing? Table 6.1 offers an answer.

Isn't it interesting that in the case of the receptive skills, we can observe neither the process of performing nor a product? I can hear your argument already: "But I can *see* that she's listening because she's nodding her head and frowning and smiling and asking relevant questions." Well, you're not observing the listening performance; you're observing the *result* of the listening. You can no more observe listening (or reading) than you can see the wind blowing. The process of

Table 6.1. Observable performance of the four skills

	Can the teacher directly observe . . .	
	the process?	the product?
Listening	No	No
Speaking	Yes	No*
Reading	No	No
Writing	Yes	Yes

*Except in the case of an audio or video recording that preserves the output.

the listening performance itself is the *invisible, inaudible* process of interna izing meaning from the auditory signals being transmitted to the ear and brair Or you may argue that the product of listening is a spoken or written respons from the student that indicates correct (or incorrect) auditory processing. Agair the product of listening and reading is not the spoken or written response. Th product is within the structure of the brain, and until teachers carry with ther little portable MRI scanners to detect meaningful intake, it is impossible t observe the product. You observe only the result of the meaningful input in th form of spoken or written output, just as you observe the result of the wind b noticing trees waving back and forth.

The productive skills of speaking and writing allow us to hear and see th process as it is performed. Writing gives a permanent product in the form of written piece. But unless you have recorded speech, there is no *permanent* obser able product for speaking performance because all those words you just heard hav vanished from your perception and (you hope) have been transformed into mear ingful intake somewhere in your brain.

Receptive skills, then, are clearly the more enigmatic of the two modes of pe formance. You cannot observe the actual act of listening or reading, nor can you se or hear an actual product! You can observe learners only *while* they are listening c reading. The upshot is that all assessment of listening and reading must be made o the basis of observing the test-taker's speaking or writing (or nonverbal response and not on the listening or reading itself. So, all assessment of receptive perfo mance must be made by inference!

How discouraging, right? Well, not necessarily. We have developed reasonabl good assessment tasks to make the necessary jump, through the process of infe ence, from unobservable reception to a conclusion about comprehension comp tence. And all this is a good reminder of the importance not just of triangulation bu of the potential fragility of the assessment of comprehension ability. The actual pe formance is made "behind the scenes," and those of us who propose to make rel able assessments of receptive performance need to be on our guard.

THE IMPORTANCE OF LISTENING

Listening has often played second fiddle to its counterpart, speaking. In the standardized testing industry, a number of separate oral production tests are available (Test of Spoken English, Oral Proficiency Inventory, and PhonePass®, to name several that are described Chapter 7 of this book), but it is rare to find just a listening test. One reason for this emphasis is that listening is often implied as a component of speaking. How could you speak a language without also listening? In addition, the overtly observable nature of speaking renders it more empirically measurable then listening. But perhaps a deeper cause lies in universal biases toward speaking. A good speaker is often (unwisely) valued more highly than a good listener. To determine if someone is a proficient user of a language, people customarily ask, "Do you speak Spanish?" People rarely ask, "Do you *understand* and speak Spanish?"

Every teacher of language knows that one's oral production ability—other than monologues, speeches, reading aloud, and the like—is only as good as one's listening comprehension ability. But of even further impact is the likelihood that *input* in the aural-oral mode accounts for a large proportion of successful language acquisition. In a typical day, we do measurably more listening than speaking (with the exception of one or two of your friends who may be nonstop chatterboxes!). Whether in the workplace, educational, or home contexts, aural comprehension far outstrips oral production in quantifiable terms of time, number of words, effort, and attention.

We therefore need to pay close attention to listening as a mode of performance for assessment in the classroom. In this chapter, we will begin with basic principles and types of listening, then move to a survey of tasks that can be used to assess listening. (For a review of issues in teaching listening, you may want to read Chapter 16 of *TBP.*)

BASIC TYPES OF LISTENING

As with all effective tests, designing appropriate assessment tasks in listening begins with the specification of objectives, or criteria. Those objectives may be classified in terms of several types of listening performance. Think about what you do when you listen. Literally in nanoseconds, the following processes flash through your brain:

1. You recognize speech sounds and hold a temporary "imprint" of them in short-term memory.
2. You simultaneously determine the type of speech event (monologue, interpersonal dialogue, transactional dialogue) that is being processed and attend to its context (who the speaker is, location, purpose) and the content of the message.
3. You use (bottom-up) linguistic decoding skills and/or (top-down) background schemata to bring a plausible interpretation to the message, and assign a *literal* and *intended meaning* to the utterance.

4. In most cases (except for repetition tasks, which involve short-term memory only), you delete the exact linguistic form in which the message was originally received in favor of conceptually retaining important or relevant information in long-term memory.

Each of these stages represents a potential assessment objective:

- comprehending of surface structure elements such as phonemes, words, intonation, or a grammatical category
- understanding of pragmatic context
- determining meaning of auditory input
- developing the gist, a global or comprehensive understanding

From these stages we can derive four commonly identified types of listening performance, each of which comprises a category within which to consider assessment tasks and procedures.

1. *Intensive.* Listening for perception of the components (phonemes, words, intonation, discourse markers, etc.) of a larger stretch of language.
2. *Responsive.* Listening to a relatively short stretch of language (a greeting, question, command, comprehension check, etc.) in order to make an equally short response.
3. *Selective.* Processing stretches of discourse such as short monologues for several minutes in order to "scan" for certain information. The purpose of such performance is not necessarily to look for global or general meanings, but to be able to comprehend designated information in a context of longer stretches of spoken language (such as classroom directions from a teacher, TV or radio news items, or stories). Assessment tasks in selective listening could ask students, for example, to listen for names, numbers, a grammatical category, directions (in a map exercise), or certain facts and events.
4. *Extensive.* Listening to develop a top-down, global understanding of spoken language. Extensive performance ranges from listening to lengthy lectures to listening to a conversation and deriving a comprehensive message or purpose. Listening for the gist, for the main idea, and making inferences are all part of extensive listening.

For full comprehension, test-takers may at the extensive level need to invoke interactive skills (perhaps note-taking, questioning, discussion): listening that includes all four of the above types as test-takers actively participate in discussions, debates, conversations, role plays, and pair and group work. Their listening performance must be intricately integrated with speaking (and perhaps other skills) in the authentic give-and-take of communicative interchange. (Assessment of interactive skills will be embedded in Chapter 7.)

MICRO- AND MACROSKILLS OF LISTENING

A useful way of synthesizing the above two lists is to consider a finite number of micro- and macroskills implied in the performance of listening comprehension. Richards' (1983) list of microskills has proven useful in the domain of specifying objectives for learning and may be even more useful in forcing test makers to carefully identify specific assessment objectives. In the following box, the skills are subdivided into what I prefer to think of as microskills (attending to the smaller bits and chunks of language, in more of a bottom-up process) and macroskills (focusing on the larger elements involved in a top-down approach to a listening task). The micro- and macroskills provide 17 different objectives to assess in listening.

Micro- and macroskills of listening (adapted from Richards, 1983)

Microskills

1. Discriminate among the distinctive sounds of English.
2. Retain chunks of language of different lengths in short-term memory.
3. Recognize English stress patterns, words in stressed and unstressed positions, rhythmic structure, intonation contours, and their role in signaling information.
4. Recognize reduced forms of words.
5. Distinguish word boundaries, recognize a core of words, and interpret word order patterns and their significance.
6. Process speech at different rates of delivery.
7. Process speech containing pauses, errors, corrections, and other performance variables.
8. Recognize grammatical word classes (nouns, verbs, etc.), systems (e.g., tense, agreement, pluralization), patterns, rules, and elliptical forms.
9. Detect sentence constituents and distinguish between major and minor constituents.
10. Recognize that a particular meaning may be expressed in different grammatical forms.
11. Recognize cohesive devices in spoken discourse.

Macroskills

12. Recognize the communicative functions of utterances, according to situations, participants, goals.
13. Infer situations, participants, goals using real-world knowledge.
14. From events, ideas, and so on, described, predict outcomes, infer links and connections between events, deduce causes and effects, and detect such relations as main idea, supporting idea, new information, given information, generalization, and exemplification.

15. Distinguish between literal and implied meanings.
16. Use facial, kinesic, body language, and other nonverbal clues to decipher meanings.
17. Develop and use a battery of listening strategies, such as detecting key words, guessing the meaning of words from context, appealing for help, and signaling comprehension or lack thereof.

Implied in the taxonomy above is a notion of what makes many aspects of listening difficult, or why listening is not simply a linear process of recording strings of language as they are transmitted into our brains. Developing a sense of which aspects of listening performance are predictably difficult will help you to challenge your students appropriately and to assign weights to items. Consider the following list of what makes listening difficult (adapted from Richards, 1983; Ur, 1984; Dunkel, 1991):

1. *Clustering:* attending to appropriate "chunks" of language—phrases, clauses, constituents
2. *Redundancy:* recognizing the kinds of repetitions, rephrasing, elaborations, and insertions that unrehearsed spoken language often contains, and benefiting from that recognition
3. *Reduced forms:* understanding the reduced forms that may not have been a part of an English learner's past learning experiences in classes where only formal "textbook" language has been presented
4. *Performance variables:* being able to "weed out" hesitations, false starts, pauses, and corrections in natural speech
5. *Colloquial language:* comprehending idioms, slang, reduced forms, shared cultural knowledge
6. *Rate of delivery:* keeping up with the speed of delivery, processing automatically as the speaker continues
7. *Stress, rhythm, and intonation:* correctly understanding prosodic elements of spoken language, which is almost always much more difficult than understanding the smaller phonological bits and pieces
8. *Interaction:* managing the interactive flow of language from listening to speaking to listening, etc.

DESIGNING ASSESSMENT TASKS: INTENSIVE LISTENING

Once you have determined objectives, your next step is to design the tasks, including making decisions about how you will elicit performance and how you will expect the test-taker to respond. We will look at tasks that range from intensive listening performance, such as minimal phonemic pair recognition, to extensive comprehension of language in communicative contexts. The focus in this section is on the microskills of intensive listening.

ecognizing Phonological and Morphological Elements

A typical form of intensive listening at this level is the assessment of recognition of phonological and morphological elements of language. A classic test task gives a spoken stimulus and asks test-takers to identify the stimulus from two or more choices, as in the following two examples:

Phonemic pair, consonants

Test-takers hear:	He's from California.
Test-takers read:	(a) He's from California. (b) She's from California.

Phonemic pair, vowels

Test-takers hear:	Is he living?
Test-takers read:	(a) Is he leaving? (b) Is he living?

In both cases above, minimal phonemic distinctions are the target. If you are testing recognition of morphology, you can use the same format:

Morphological pair, -ed ending

Test-takers hear:	I missed you very much.
Test-takers read:	(a) I missed you very much. (b) I miss you very much.

Hearing the past tense morpheme in this sentence challenges even advanced learners, especially if no context is provided. Stressed and unstressed words may also be tested with the same rubric. In the following example, the reduced form (contraction) of *can not* is tested:

Stress pattern in can't

Test-takers hear:	My girlfriend can't go to the party.
Test-takers read:	(a) My girlfriend can't go to the party. (b) My girlfriend can go to the party.

Because they are decontextualized, these kinds of tasks leave something to b desired in their authenticity. But they are a step better than items that simply pr vide a one-word stimulus:

One-word stimulus

Test-takers hear:	vine
Test-takers read:	(a) vine (b) wine

Paraphrase Recognition

The next step up on the scale of listening comprehension microskills is words, phrase and sentences, which are frequently assessed by providing a stimulus sentence an asking the test-taker to choose the correct paraphrase from a number of choices.

Sentence paraphrase

Test-takers hear:	Hellow, my name's Keiko. I come from Japan.
Test-takers read:	(a) Keiko is comfortable in Japan. (b) Keiko wants to come to Japan. (c) Keiko is Japanese. (d) Keiko likes Japan.

In the above item, the idiomatic *come from* is the phrase being tested. To add a litt context, a conversation can be the stimulus task to which test-takers must respon with the correct paraphrase:

Dialogue paraphrase

Test-takers hear:	Man:	Hi, Maria, my name's George.
	Woman:	Nice to meet you, George. Are you American?
	Man:	No, I'm Canadian.
Test-takers read:		(a) George lives in the United States. (b) George is American. (c) George comes from Canada. (d) Maria is Canadian.

Here, the criterion is recognition of the adjective form used to indicate country origin: Canadian, American, Brazilian, Italian, etc.

DESIGNING ASSESSMENT TASKS: RESPONSIVE LISTENING

A question-and-answer format can provide some interactivity in these lower-end listening tasks. The test-taker's response is the appropriate answer to a question.

Appropriate response to a question

Test-takers hear:	How much time did you take to do your homework?
Test-takers read:	(a) In about an hour. (b) About an hour. (c) About $10. (d) Yes, I did.

The objective of this item is recognition of the *wh*-question *how much* and its appropriate response. Distractors are chosen to represent common learner errors: (a) responding to *how much* vs. *how much longer;* (c) confusing *how much* in reference to time vs. the more frequent reference to money; (d) confusing a *wh*-question with a *yes/no* question.

None of the tasks so far discussed have to be framed in a multiple-choice format. They can be offered in a more open-ended framework in which test-takers write or speak the response. The above item would then look like this:

Open-ended response to a question

Test-takers hear:	How much time did you take to do your homework?
Test-takers write or speak:	_____.

If open-ended response formats gain a small amount of authenticity and creativity, they of course suffer some in their practicality, as teachers must then read students' responses and judge their appropriateness, which takes time.

DESIGNING ASSESSMENT TASKS: SELECTIVE LISTENING

A third type of listening performance is selective listening, in which the test-taker listens to a limited quantity of aural input and must discern within it some specific information. A number of techniques have been used that require selective listening.

Listening Cloze

Listening cloze tasks (sometimes called cloze dictations or partial dictations) require the test-taker to listen to a story, monologue, or conversation and simultaneously

read the written text in which selected words or phrases have been deleted. Cloze procedure is most commonly associated with reading only (see Chapter 9). In its generic form, the test consists of a passage in which every *n*th word (typically every seventh word) is deleted and the test-taker is asked to supply an appropriate word. In a listening cloze task, test-takers see a transcript of the passage that they are listening to and fill in the blanks with the words or phrases that they hear.

One potential weakness of listening cloze techniques is that they may simply become reading comprehension tasks. Test-takers who are asked to listen to a story with periodic deletions in the written version may not need to listen at all, yet may still be able to respond with the appropriate word or phrase. You can guard against this eventuality if the blanks are items with high information load that cannot be easily predicted simply by reading the passage. In the example below (adapted from Bailey, 1998, p. 16), such a shortcoming was avoided by focusing only on the criterion of numbers. Test-takers hear an announcement from an airline agent and see the transcript with the underlined words deleted:

Listening cloze

Test-takers hear:

Ladies and gentlemen, I now have some connecting gate information for those of you making connections to other flights out of San Francisco.

Flight *seven-oh-six* to Portland will depart from gate *seventy-three* at *nine-thirty* P.M.
Flight *ten-forty-five* to Reno will depart at *nine-fifty* P.M. from gate *seventeen*.
Flight *four-forty* to Monterey will depart at *nine-thirty-five* P.M. from gate *sixty*.
And flight *sixteen-oh-three* to Sacramento will depart from gate *nineteen* at *ten-fifteen* P.M.

Test-takers write the missing words or phrases in the blanks.

Other listening cloze tasks may focus on a grammatical category such as verb tenses, articles, two-word verbs, prepositions, or transition words/phrases. Notice two important structural differences between listening cloze tasks and standard reading cloze. In a listening cloze, deletions are governed by the objective of the test, not by mathematical deletion of every *n*th word; and more than one word may be deleted, as in the above example.

Listening cloze tasks should normally use an **exact word** method of scoring, in which you accept as a correct response only the actual word or phrase that was spoken and consider other **appropriate words** as incorrect. (See Chapter 8 for further discussion of these two methods.) Such stringency is warranted; your objective is, after all, to test listening comprehension, not grammatical or lexical expectancies.

Information Transfer

Selective listening can also be assessed through an **information transfer** technique in which aurally processed information must be transferred to a visual representation, such as labeling a diagram, identifying an element in a picture, completing a form, or showing routes on a map.

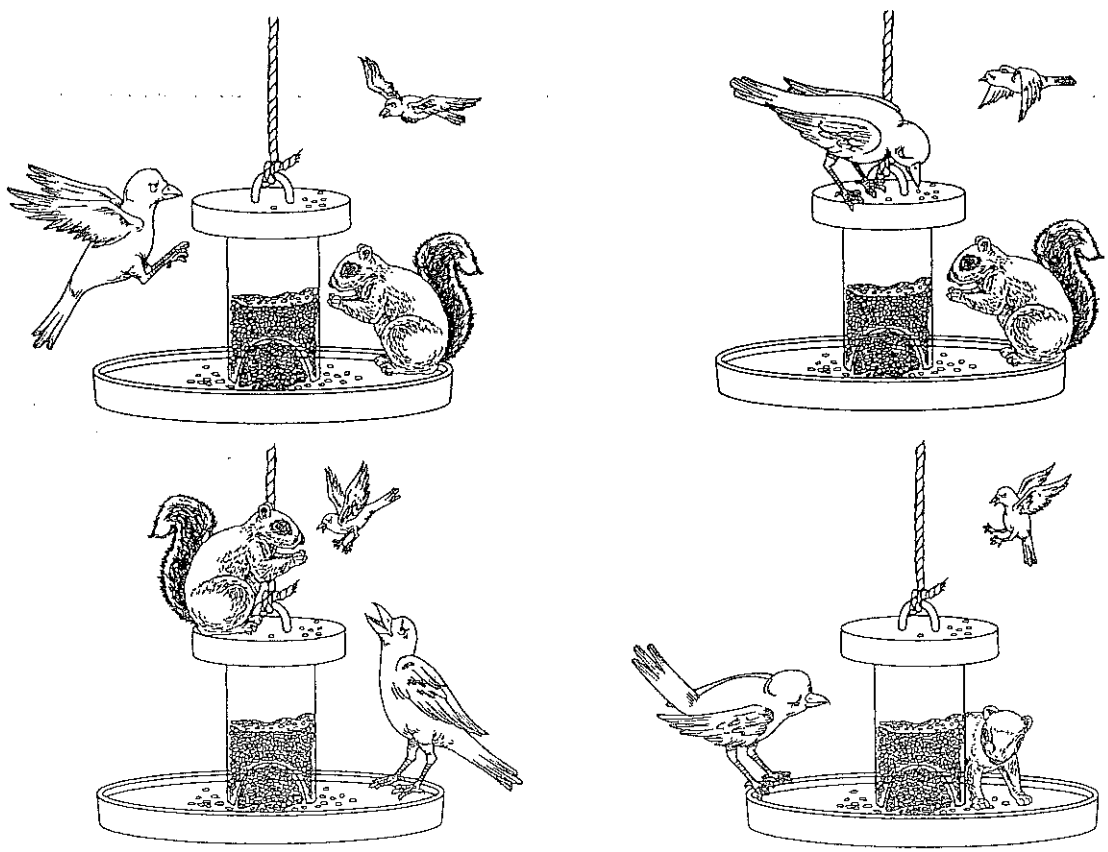
At the lower end of the scale of linguistic complexity, simple **picture-cued** items are sometimes efficient rubrics for assessing certain selected information. Consider the following item:

Information transfer: multiple-picture–cued selection

Test-takers hear:

Choose the correct picture. In my back yard I have a bird feeder. Yesterday, there were two birds and a squirrel fighting for the last few seeds in the bird feeder. The squirrel was on top of the bird feeder while the larger bird sat at the bottom of the feeder screeching at the squirrel. The smaller bird was flying around the squirrel, trying to scare it away.

Test-takers see:



The preceding example illustrates the need for test-takers to focus on just the relevant information. The objective of this task is to test prepositions and prepositional phrases of location (*at the bottom, on top of, around,* along with *larger, smaller*), so other words and phrases such as *back yard, yesterday, last few seeds,* and *scare away* are supplied only as context and need not be tested. (The task also presupposes, of course, that test-takers are able to identify the difference between a bird and a squirrel!)

In another genre of picture-cued tasks, a number of people and/or actions are presented in one picture, such as a group of people at a party. Assuming that all the items, people, and actions are clearly depicted and understood by the test-taker, assessment may take the form of

- questions: "Is the tall man near the door talking to a short woman?"
- true/false: "The woman wearing a red skirt is watching TV."
- identification: "Point to the person who is standing behind the lamp." "Draw a circle around the person to the left of the couch."

In a third picture-cued option used by the Test of English for International Communication (TOEIC®), one single photograph is presented to the test-taker, who then hears four different statements and must choose one of the four to describe the photograph. Here is an example.

Information transfer: single-picture–cued verbal multiple-choice

Test-takers see:	a photograph of a woman in a laboratory setting, with no glasses on, squinting through a microscope with her right eye, and with her left eye closed.
Test-takers hear:	(a) She's speaking into a microphone. (b) She's putting on her glasses. (c) She has both eyes open. (d) She's using a microscope.

Information transfer tasks may reflect greater authenticity by using charts, maps, grids, timetables, and other artifacts of daily life. In the example below, test-takers hear a student's daily schedule, and the task is to fill in the partially completed weekly calendar.

Information transfer: chart-filling

Test-takers hear:

Now you will hear information about Lucy's daily schedule. The information will be given twice. The first time just listen carefully. The second time, there will be a pause after each sentence. Fill in Lucy's blank daily schedule with the correct information. The example has already been filled in.

You will hear: Lucy gets up at eight o'clock every morning except on weekends.

You will fill in the schedule to provide the information.

Now listen to the information about Lucy's schedule. Remember, you will first hear all the sentences; then you will hear each sentence separately with time to fill in your chart.

Lucy gets up at 8:00 every morning except on weekends. She has English on Monday, Wednesday, and Friday at ten o'clock. She has History on Tuesdays and Thursdays at two o'clock. She takes Chemistry on Monday from two o'clock to six o'clock. She plays tennis on weekends at four o'clock. She eats lunch at twelve o'clock every day except Saturday and Sunday.

Now listen a second time. There will be a pause after each sentence to give you time to fill in the chart. (Lucy's schedule is repeated with a pause after each sentence).

Test-takers see the following weekly calendar grid:

	Monday	Tuesday	Wednesday	Thursday	Friday	Weekends
8:00	get up	get up	get up	get up	get up	
10:00						
12:00						
2:00						
4:00						
6:00						

Such chart-filling tasks are good examples of aural **scanning** strategies. A listener must discern from a number of pieces of information which pieces are relevant. In the above example, virtually all of the stimuli are relevant, and very few words can be ignored. In other tasks, however, much more information might be presented than is needed (as in the birdfeeder item on page 127), forcing the test-taker to select the correct bits and pieces necessary to complete a task.

Chart-filling tasks increase in difficulty as the linguistic stimulus material becomes more complex. In one task described by Ur (1984, pp. 108–112), test-takers listen to a very long description of animals in various cages in a zoo. While they listen, they can look at a map of the layout of the zoo with unlabeled cages. Their task is to fill in the correct animal in each cage, but the complexity of the language used to describe the positions of cages and their inhabitants is very challenging. Similarly, Hughes (1989, p. 138) described a map-marking task in which test-takers must process around 250 words of colloquial language in order to complete the tasks of identifying names, positions, and directions in a car accident scenario on a city street.

Sentence Repetition

The task of simply repeating a sentence or a partial sentence, or sentence repetition, is also used as an assessment of listening comprehension. As in a dictation (discussed below), the test-taker must retain a stretch of language long enough to reproduce it, and then must respond with an oral repetition of that stimulus. Incorrect listening comprehension, whether at the phonemic or discourse level, may be manifested in the correctness of the repetition. A miscue in repetition is scored as a miscue in listening. In the case of somewhat longer sentences, one could argue that the ability to recognize and retain chunks of language as well as thread of meaning might be assessed through repetition. In Chapter 7, we will look closely at PhonePass, a commercially produced test that relies largely on sentence repetition to assess both oral production and listening comprehension.

Sentence repetition is far from a flawless listening assessment task. Buck (2001, p. 79) noted that such tasks "are not just tests of listening, but tests of general oral skills." Further, this task may test only recognition of sounds, and it can easily be contaminated by lack of short-term memory ability, thus invalidating it as an assessment of comprehension alone. And the teacher may never be able to distinguish a listening comprehension error from an oral production error. Therefore, sentence repetition tasks should be used with caution.

DESIGNING ASSESSMENT TASKS: EXTENSIVE LISTENING

Drawing a clear distinction between any two of the categories of listening referred to here is problematic, but perhaps the fuzziest division is between selective and extensive listening. As we gradually move along the continuum from smaller to larger stretches of language, and from micro- to macroskills of listening, the probability of using more extensive listening tasks increases. Some important questions about designing assessments at this level emerge.

1. Can listening performance be distinguished from cognitive processing factors such as memory, associations, storage, and recall?
2. As assessment procedures become more communicative, does the task take into account test-takers' ability to use grammatical expectancies, lexical collocations, semantic interpretations, and pragmatic competence?
3. Are test tasks themselves correspondingly content valid and authentic—that is, do they mirror real-world language and context?
4. As assessment tasks become more and more open-ended, they more closely resemble pedagogical tasks, which leads one to ask what the difference is between assessment and teaching tasks. The answer is *scoring:* the former imply specified scoring procedures, while the latter do not.

We will try to address these questions as we look at a number of extensive or quasi-extensive listening comprehension tasks.

Dictation

Dictation is a widely researched genre of assessing listening comprehension. In a dictation, test-takers hear a passage, typically of 50 to 100 words, recited three times: first, at normal speed; then, with long pauses between phrases or natural word groups, during which time test-takers write down what they have just heard; and finally, at normal speed once more so they can check their work and proofread. Here is a sample dictation at the intermediate level of English.

Dictation

First reading (natural speed, no pauses, test-takers listen for gist):

The state of California has many geographical areas. On the western side is the Pacific Ocean with its beaches and sea life. The central part of the state is a large fertile valley. The southeast has a hot desert, and north and west have beautiful mountains and forests. Southern California is a large urban area populated by millions of people.

Second reading (slowed speed, pause at each // break, test-takers write):

The state of California // has many geographical areas. // On the western side // is the Pacific Ocean // with its beaches and sea life. // The central part of the state // is a large fertile valley. // The southeast has a hot desert, // and north and west // have beautiful mountains and forests. // Southern California // is a large urban area // populated by millions of people.

Third reading (natural speed, test-takers check their work).

Dictations have been used as assessment tools for decades. Some readers still cringe at the thought of having to render a correctly spelled, verbatim version of a paragraph or story recited by the teacher. Until research on integrative testing was published (see Oller, 1971), dictations were thought to be not much more than glorified spelling tests. However, the required integration of listening and writing in a dictation, along with its presupposed knowledge of grammatical and discourse expectancies, brought this technique back into vogue. Hughes (1989), Cohen (1994), Bailey (1998), and Buck (2001) all defend the plausibility of dictation as an integrative test that requires some sophistication in the language in order to process and write down all segments correctly. Thus, I include dictation here under the rubric of extensive tasks, although I am more comfortable with labeling it quasi-extensive.

The difficulty of a dictation task can be easily manipulated by the length of the word groups (or **bursts**, as they are technically called), the length of the pauses, the speed at which the text is read, and the complexity of the discourse, grammar, and vocabulary used in the passage.

Scoring is another matter. Depending on your context and purpose in administering a dictation, you will need to decide on scoring criteria for several possible kinds of errors:

- spelling error only, but the word appears to have been heard correctly
- spelling and/or obvious misrepresentation of a word, illegible word
- grammatical error (For example, test-taker hears *I can't do it,* writes *I can do it.*)
- skipped word or phrase
- permutation of words
- additional words not in the original
- replacement of a word with an appropriate synonym

Determining the weight of each of these errors is a highly idiosyncratic choice; specialists disagree almost more than they agree on the importance of the above categories. They do agree (Buck, 2001) that a dictation is not a spelling test, and that the first item in the list above should not be considered an error. They also suggest that point systems be kept simple (for maintaining practicality and reliability) and that a deductible scoring method, in which points are subtracted from a hypothetical total, is usually effective.

Dictation seems to provide a reasonably valid method for integrating listening and writing skills and for tapping into the cohesive elements of language implied in short passages. However, a word of caution lest you assume that dictation provides a quick and easy method of assessing extensive listening comprehension. If the bursts in a dictation are relatively long (more than five-word segments), this method places a certain amount of load on memory and processing of meaning (Buck, 2001, p. 78). But only a moderate degree of cognitive processing is required, and claiming that dictation fully assesses the ability to comprehend pragmatic or illocutionary elements of language, context, inference, or semantics may be going too far. Finally, one can easily question the authenticity of dictation: it is rare in the real world for people to write down more than a few chunks of information (addresses, phone numbers, grocery lists, directions, for example) at a time.

Despite these disadvantages, the practicality of the administration of dictations, a moderate degree of reliability in a well-established scoring system, and a strong correspondence to other language abilities speaks well for the inclusion of dictation among the possibilities for assessing extensive (or quasi-extensive) listening comprehension.

Communicative Stimulus-Response Tasks

Another—and more authentic—example of extensive listening is found in a popular genre of assessment task in which the test-taker is presented with a stimulus monologue or conversation and then is asked to respond to a set of comprehension questions. Such tasks (as you saw in Chapter 4 in the discussion of standardized testing) are commonly used in commercially produced proficiency tests. The monologues, lectures, and brief conversations used in such tasks are sometimes a little contrived,

and certainly the subsequent multiple-choice questions don't mirror communicative, real-life situations. But with some care and creativity, one can create reasonably authentic stimuli, and in some rare cases the response mode (as shown in one example below) actually approaches complete authenticity. Here is a typical example of such a task.

Dialogue and multiple-choice comprehension items

Test-takers hear:

Directions: Now you will hear a conversation between Lynn and her doctor. You will hear the conversation two times. After you hear the conversation the second time, choose the correct answer for questions 11–15 below. Mark your answers on the answer sheet provided.

Doctor:	Good morning, Lynn. What's the problem?
Lynn:	Well, you see, I have a terrible headache, my nose is running, and I'm really dizzy.
Doctor:	Okay. Anything else?
Lynn:	I've been coughing, I think I have a fever, and my stomach aches.
Doctor:	I see. When did this start?
Lynn:	Well, let's see, I went to the lake last weekend, and after I returned home I started sneezing.
Doctor:	Hmm. You must have the flu. You should get lots of rest, drink hot beverages, and stay warm. Do you follow me?
Lynn:	Well, uh, yeah, but . . . shouldn't I take some medicine?
Doctor:	Sleep and rest are as good as medicine when you have the flu.
Lynn:	Okay, thanks, Dr. Brown.

Test-takers read:

11. What is Lynn's problem?
(A) She feels horrible.
(B) She ran too fast at the lake.
(C) She's been drinking too many hot beverages.

12. When did Lynn's problem start?
(A) When she saw her doctor.
(B) Before she went to the lake.
(C) After she came home from the lake.

13. The doctor said that Lynn _____.
(A) flew to the lake last weekend
(B) must not get the flu
(C) probably has the flu

14. The doctor told Lynn _____.
    (A) to rest
    (B) to follow him
    (C) to take some medicine
15. According to Dr. Brown, sleep and rest are _____ medicine when you have the flu.
    (A) more effective than
    (B) as effective as
    (C) less effective than

Does this meet the criterion of authenticity? If you want to be painfully fussy, you might object that it is rare in the real world to eavesdrop on someone else's doctor–patient conversation. Nevertheless, the conversation itself is relatively authentic; we all have doctor–patient exchanges like this. Equally authentic, if you add a grain of salt, are monologues, lecturettes, and news stories, all of which are commonly utilized as listening stimuli to be followed by comprehension questions aimed at assessing certain objectives that are built into the stimulus.

Is the task itself (of responding to multiple-choice questions) authentic? It's plausible to assert that *any task* of this kind following a one-way listening to a conversation is artificial: we simply don't often encounter little quizzes about conversations we've heard (unless it's your parent, spouse, or best friend who wants to get in on the latest gossip!). The questions posed above, with the possible exception of **#14**, are unlikely to appear in a lifetime of doctor visits. Yet the ability to respond correctly to such items can be construct validated as an appropriate measure of field-independent listening skills: the ability to remember certain details from a conversation. (As an aside here, many highly proficient native speakers of English might miss some of the above questions if they heard the conversation only once and if they had no visual access to the items until after the conversation was done!)

To compensate for the potential inauthenticity of post-stimulus comprehension questions, you might, with a little creativity, be able to find contexts where questions that probe understanding are more appropriate. Consider the following situation:

Dialogue and authentic questions on details

Test-takers hear:
You will hear a conversation between a detective and a man. The tape will play the conversation twice. After you hear the conversation a second time, choose the correct answers on your test sheet.

Detective:	Where were you last night at eleven P.M., the time of the murder?
Man:	Uh, let's see; well, I was just starting to see a movie.
Detective:	Did you go alone?
Man:	No, uh, well, I was with my friend, uh, Bill. Yeah, I was with Bill.

Detective: What did you do after that?
Man: We went out to dinner, then I dropped her off at her place.
Detective: Then you went home?
Man: Yeah.
Detective: When did you get home?
Man: A little before midnight.

Test-takers read:

7. Where was the man at 11:00 P.M.?
   (A) In a restaurant.
   (B) In a theater.
   (C) At home.
8. Was he with someone?
   (A) He was alone.
   (B) He was with his wife.
   (C) He was with a friend.
9. Then what did he do?
   (A) He ate out.
   (B) He made dinner.
   (C) He went home.
10. When did he get home?
    (A) About 11:00.
    (B) Almost 12:00.
    (C) Right after the movie.
11. The man is probably lying because (name two clues):
    1. _____
    2. _____

In this case, test-takers are brought into a little scene in a crime story. The questions following are plausible questions that might be asked to review fact and fiction in the conversation. Question **#11,** of course, provides an extra shot of reality: the test-taker must name the probable lies told by the man (he referred to Bill as "her"; he saw a movie and ate dinner in the space of one hour), which requires the process of inference.

Authentic Listening Tasks

Ideally, the language assessment field would have a stockpile of listening test types that are cognitively demanding, communicative, and authentic, not to mention interactive by means of an integration with speaking. However, the nature of a test as a *sample* of performance and a set of tasks with limited time frames implies an equally limited capacity to mirror all the real-world contexts of listening performance. "There

is no such thing as a communicative test," stated Buck (2001, p. 92). "Every test requires some components of communicative language ability, and no test covers them all. Similarly, with the notion of authenticity, every task shares some characteristics with target-language tasks, and no test is completely authentic."

Beyond the rubrics of intensive, responsive, selective, and quasi-extensive communicative contexts described above, can we assess aural comprehension in a truly communicative context? Can we, at this end of the range of listening tasks, ascertain from test-takers that they have processed the main idea(s) of a lecture, the gist of a story, the pragmatics of a conversation, or the unspoken inferential data present in most authentic aural input? Can we assess a test-taker's comprehension of humor, idiom, and metaphor? The answer is a cautious yes, but not without some concessions to practicality. And the answer is a more certain yes if we take the liberty of stretching the concept of assessment to extend beyond tests and into a broader framework of alternatives. Here are some possibilities.

1. Note-taking. In the academic world, classroom lectures by professors are common features of a non-native English-user's experience. One form of a midterm examination at the American Language Institute at San Francisco State University (Kahn, 2002) uses a 15-minute lecture as a stimulus. One among several response formats includes note-taking by the test-takers. These notes are evaluated by the teacher on a 30-point system, as follows:

Scoring system for lecture notes

0–15 points
Visual representation: Are your notes clear and easy to read? Can you easily find and retrieve information from them? Do you use the space on the paper to visually represent ideas? Do you use indentation, headers, numbers, etc.?

0–10 points
Accuracy: Do you accurately indicate main ideas from lectures? Do you note important details and supporting information and examples? Do you leave out unimportant information and tangents?

0–5 points
Symbols and abbreviations: Do you use symbols and abbreviations as much as possible to save time? Do you avoid writing out whole words, and do you avoid writing down every single word the lecturer says?

The process of scoring is time consuming (a loss of practicality), and because of the subjectivity of the point system, it lacks some reliability. But the gain is in offering students an authentic task that mirrors exactly what they have been focusing on in the classroom. The notes become an indirect but arguably valid form of assessing global listening comprehension. The task fulfills the criteria of cognitive demand, communicative language, and authenticity.

2. *Editing.* Another authentic task provides both a written and a spoken stimulus, and requires the test-taker to listen for discrepancies. Scoring achieves relatively high reliability as there are usually a small number of specific differences that must be identified. Here is the way the task proceeds.

Editing a written version of an aural stimulus

> *Test-takers read:* the written stimulus material (a news report, an email from a friend, notes from a lecture, or an editorial in a newspaper).
>
> *Test-takers hear:* a spoken version of the stimulus that deviates, in a finite number of facts or opinions, from the original written form.
>
> *Test-takers mark:* the written stimulus by circling any words, phrases, facts, or opinions that show a discrepancy between the two versions.

One potentially interesting set of stimuli for such a task is the description of a political scandal first from a newspaper with a political bias, and then from a radio broadcast from an "alternative" news station. Test-takers are not only forced to listen carefully to differences but are subtly informed about biases in the news.

3. *Interpretive tasks.* One of the intensive listening tasks described above was paraphrasing a story or conversation. An interpretive task extends the stimulus material to a longer stretch of discourse and forces the test-taker to infer a response. Potential stimuli include

- song lyrics,
- [recited] poetry,
- radio/television news reports, and
- an oral account of an experience.

Test-takers are then directed to interpret the stimulus by answering a few questions (in open-ended form). Questions might be:

- "Why was the singer feeling sad?"
- "What events might have led up to the reciting of this poem?"
- "What do you think the political activists might do next, and why?"
- "What do you think the storyteller felt about the mysterious disappearance of her necklace?"

This kind of task moves us away from what might traditionally be considered a test toward an informal assessment, or possibly even a pedagogical technique or activity. But the task conforms to certain time limitations, and the questions can be quite specific, even though they ask the test-taker to use inference. While reliable scoring may be an issue (there may be more than one correct interpretation), the authenticity of

the interaction in this task and potential washback to the student surely give it some prominence among communicative assessment procedures.

4. *Retelling.* In a related task, test-takers listen to a story or news event and simply retell it, or summarize it, either orally (on an audiotape) or in writing. In so doing, test-takers must identify the gist, main idea, purpose, supporting points, and/or conclusion to show full comprehension. Scoring is partially predetermined by specifying a minimum number of elements that must appear in the retelling. Again reliability may suffer, and the time and effort needed to read and evaluate the response lowers practicality. Validity, cognitive processing, communicative ability, and authenticity are all well incorporated into the task.

§ § § § §

A fifth category of listening comprehension was hinted at earlier in the chapter: **interactive** listening. Because such interaction presupposes a process of *speaking* in concert with listening, the interactive nature of listening will be addressed in the next chapter. Don't forget that a significant proportion of real-world listening performance is interactive. With the exception of media input, speeches, lectures, and eavesdropping, many of our listening efforts are directed toward a two-way process of speaking and listening in face-to-face conversations.

EXERCISES

[Note: (I) Individual work; (G) Group or pair work; (C) Whole-class discussion.]

1. (C) In Table 6.1 on page 118, it is noted that one cannot actually observe listening and reading performance. Do you agree? And do you agree that there isn't even a product to observe for speaking, listening, and reading? How, then, can one infer the competence of a test-taker to speak, listen, and read a language?
2. (C) Given that we spend much more time listening than we do speaking, why are there many more tests of speaking than listening?
3. (G) Look at the list of micro- and macroskills of listening on pages 121–122. In pairs, each assigned to a different skill (or two), brainstorm some tasks that assess those skills. Present your findings to the rest of the class.
4. (G) Eight characteristics of listening that make listening "difficult" are listed on page 122. In pairs, each assigned to an assessment task itemized in this chapter, decide which of the eight factors, in order of significance, contribute to the potential difficulty of the items. Report back to the class.
5. (G) Divide the basic types of listening among groups or pairs, one type for each. Look at the sample assessment techniques provided and evaluate them

according the five principles (practicality, reliability, validity [face and content], authenticity, and washback). Present your critique to the rest of the class.

6. (G) In the same groups as in #5 above and with the same type of listening, design some other item types, different from the one(s) provided here, that assess the same type of listening performance.
7. (G) With a linguistic objective assigned to each pair or group, construct a listening cloze test for two-word verbs, verb tenses, prepositions, transition words, articles, and/or other grammatical categories.
8. (I/C) On page 131, you are reminded that dictations are considered by some assessment specialists to be integrative (requiring the integration of listening, writing, reading [proofreading], along with attendant grammatical and discourse abilities). Is this a valid claim? Justify your response.
9. (I/C) On page 136 is Buck's claim that "no test is completely authentic." Discuss the extent to which you agree or disagree with this assertion and justify your own conclusion.

OR YOUR FURTHER READING

Buck, Gary. (2001). *Assessing listening.* Cambridge: Cambridge University Press.

One of a series of very useful reference books on assessing specific skill areas published by Cambridge University Press, this one gives an overview of research and pedagogy on listening comprehension and demonstrates many different assessment procedures in common use.

Richards, Jack C. (1983). Listening comprehension: Approach, design, procedure *TESOL Quarterly, 17,* 219–239.

Even though Richards published this article in 1983, it still provides a standard backdrop for teaching listening skills. While formal assessment is not directly addressed, informal assessment is implied in its pedagogical focus on practical classroom techniques.

Mendelsohn, David J. (1998). Teaching listening. *Annual Review of Applied Linguistics, 18,* 81–101.

Mendelsohn's overview of research on teaching listening provides an excellent foundation for understanding assessment tasks. He focuses on a strategy-based approach to teaching listening and adds an annotated bibliography of professional resource books.

LISTENING COMPREHENSION TESTS

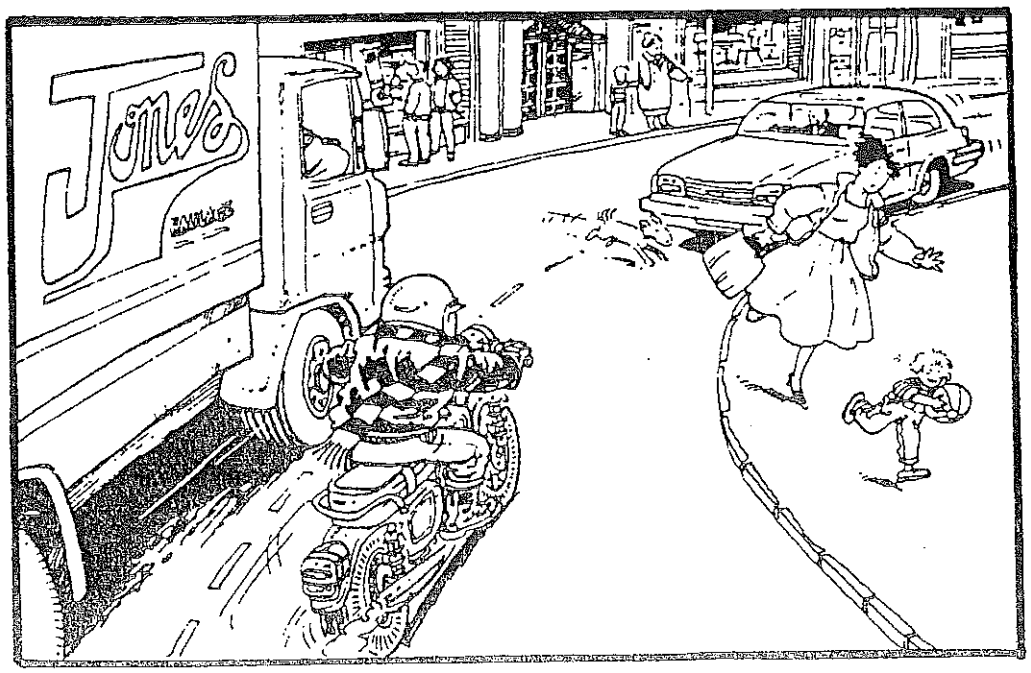
TYPE 1. The students hear a statement (usually on tape) and then Choose the best option from four written paraphrases.

Spoken : I wish you'd done it when I told you.
Written : A. I told you and you did it then.
B. I didn't tell you, but you did it then.
C. I told you, but you didn't do it then.
D. I didn't tell you and you didn't do it then.

TYPE 2. Selecting the correct response.

Spoken : Why are going home ?
Written : A. At six o'clock
B. Yes, I am
C. To help my mother
D. By bus

TYPE 3. A picture used in conjunction with spoken statements.

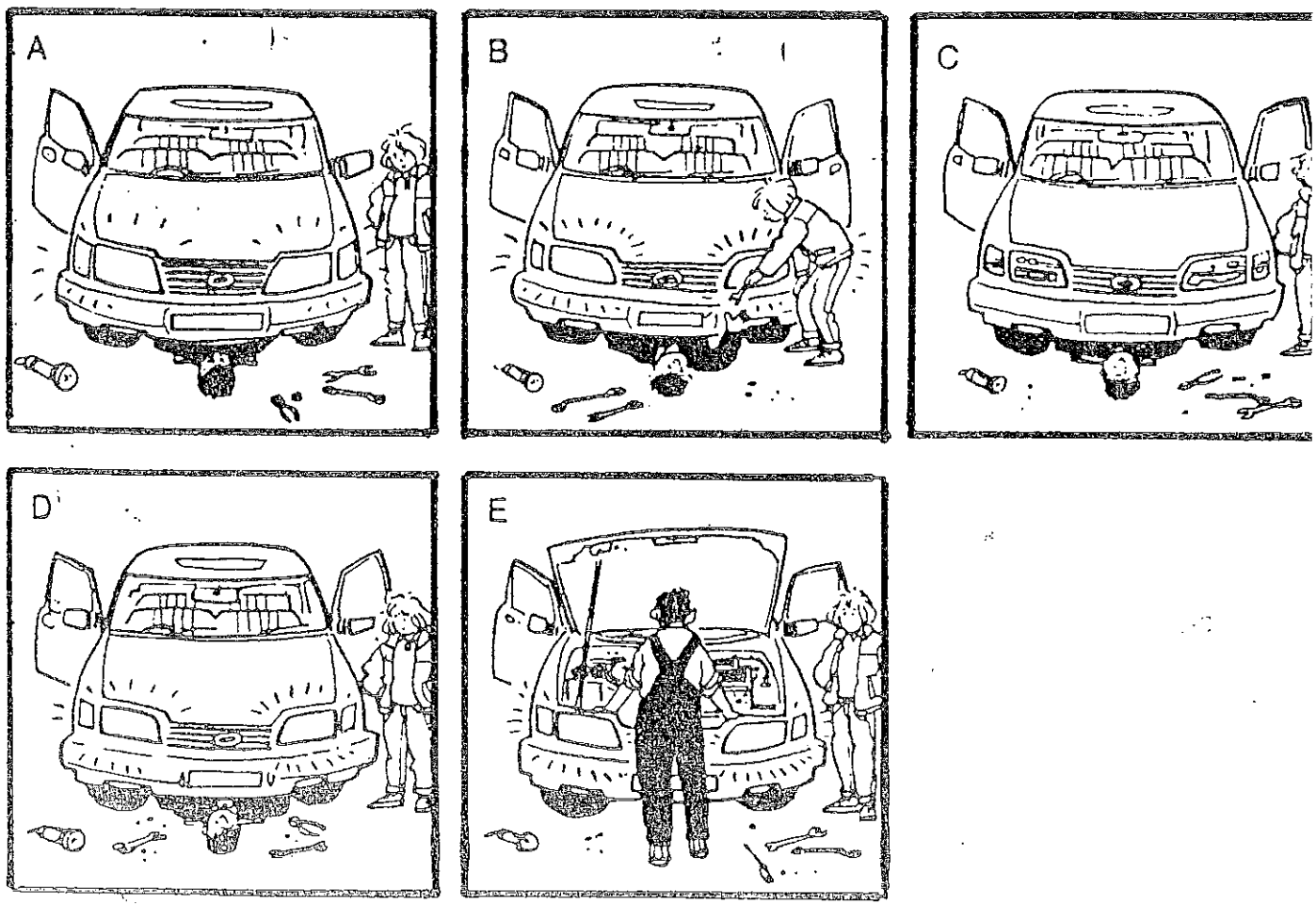


The students have to pick out the true statements and put a tick at the side of appropriate numbers. They put a cross at the side of the numbers of the false statements.

Spoken : 1. The lorry's on the left on the on the motorcyclist.
2. The car is travelling in the same direction.
3. A dog is running in front of the car.
4. A little girl's running after her mother.
5. There are a lot of cars in the street.
6. The two boysa re on the same side of the street as the little girl.

Written : 1. 3. 5.
2. 4. 6.

TYPE 4. Five or six pictures with few differences.

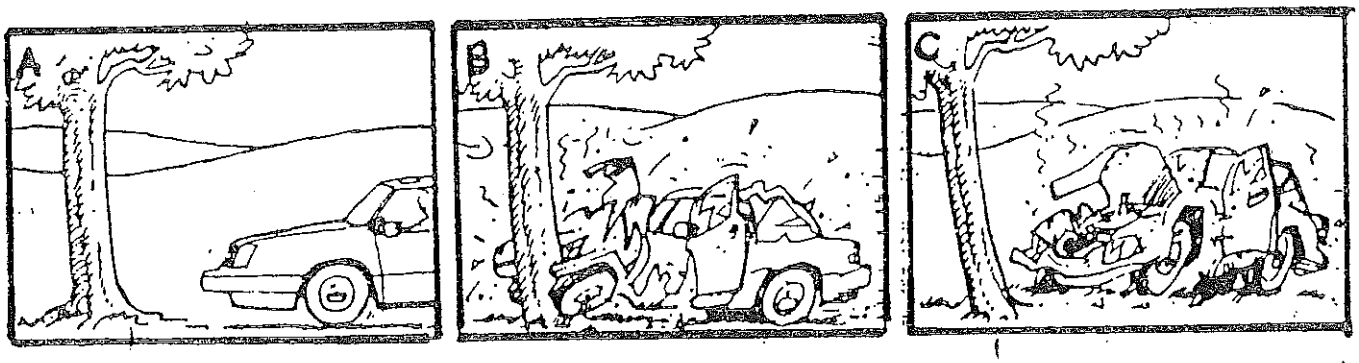


Students have five pictures in front of them. They listen to four sentences, at the end which they are required to select the appropriate picture being described.

Written : A B C D E

TYPE 5. The students see a set of three or four pictures and hear a statement, on the basis of which they have to select the most appropriate picture. In the test the students often see a total of ten or twelve such sets of pictures.

1.



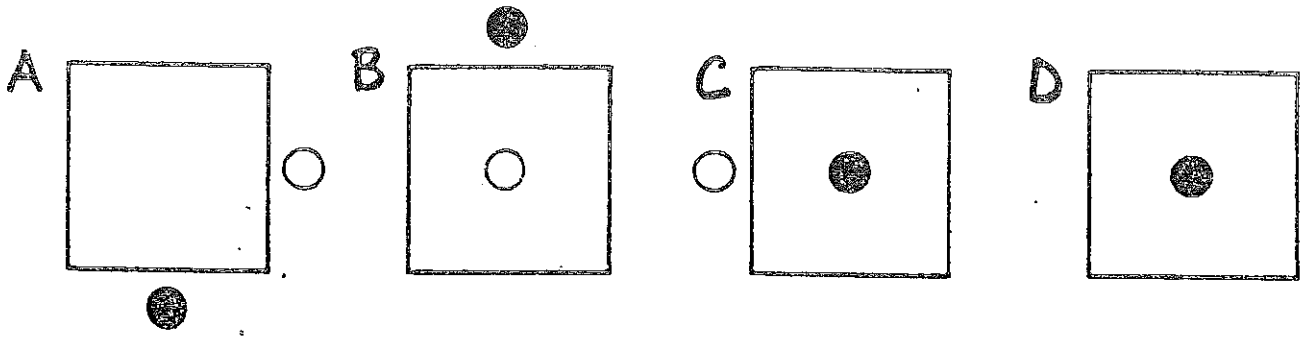
2.

The students hear :

1. The car's going to crash into a tree.
2. Danny can't run as fast as Clarie.

Written : 1. A B C
2. A B C

TYPE 6. Simple diagrams.



Look carefully at each of the four diagrams. You will hear a series of statements about each of the diagrams. Write down the approriate letter for each statement in the brackets given below each dialogue.

Spoken : 1. A : Look ! What's that inside the square ?
B : It's a white cicle.

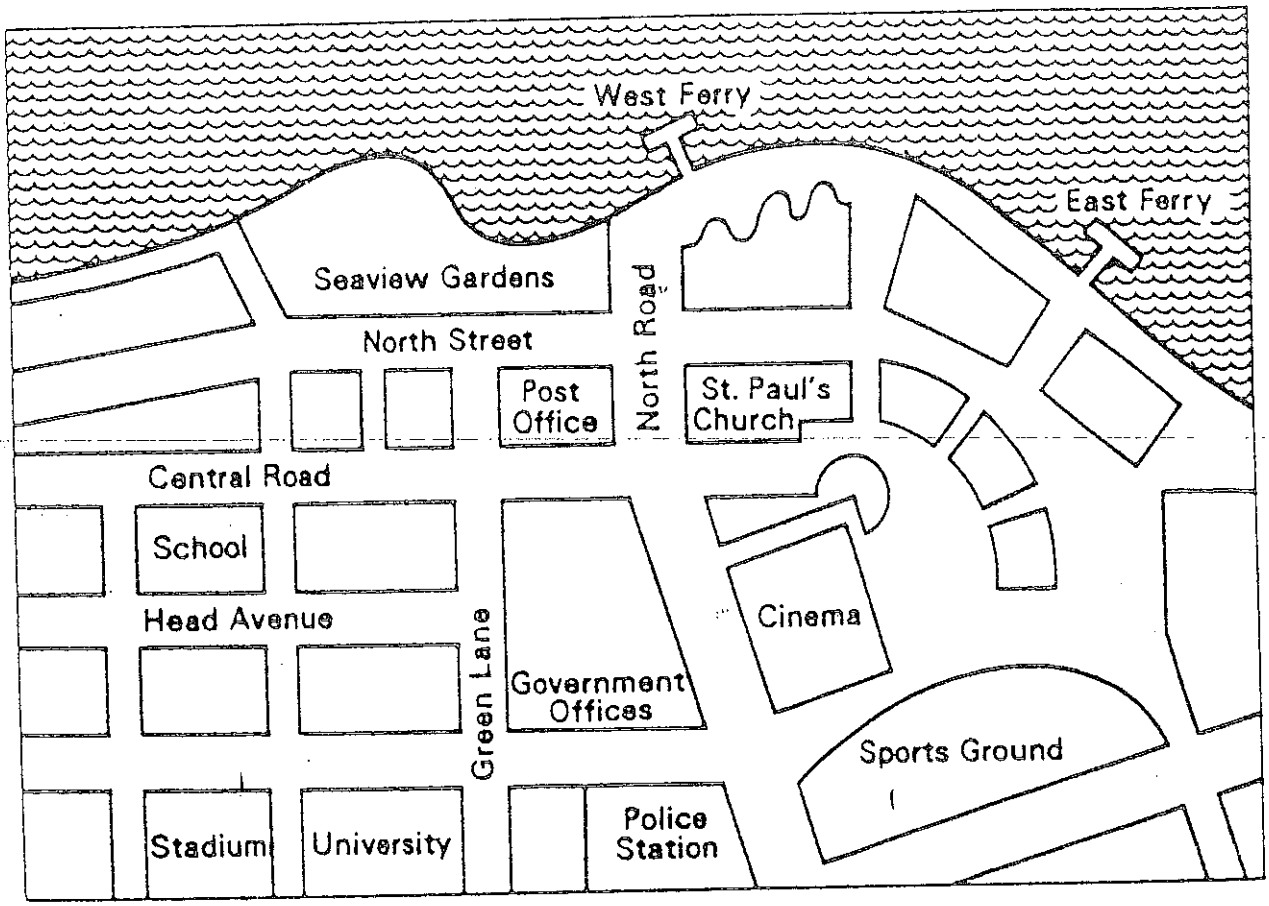
2. A : Is that a black circle ?
B : Whereabouts ?
A : Above.
B : Yes. it is. It's a black circle above the square.

3. A : Is the white circle on the left of the square ?
B : No, it's on the right of the square.

Written : 1. 2. 3.

The students should hear at least 7-8 statements.

TYPE 7. Simple instructions with a street map.



Spoken : You come out of school into Central Road and walk in the direction of Gren Lane. However, you take the left turning just before you reach Gren Lane. At the end of the street you turn right and continue until you come to the second turning right. You cross this road and you will see on your right .....
(Which building will you see ?)

TYPE 8. An incomplete picture. The students are required to add it some pieces of visual information according to certain information they are given. The students look at a simple incomplete picture and draw what they are asked to draw.

TYPE 9. Talk and lectures (intermediate and advanced).
Students listen to a short talk and select the correct statement about the talk.

Spoken : (A short talk)
Written : One of the following statemets about the talk you have just heard is correct. Put a circle round the letter next to the correct statement.

1. a. .....
   b. .....
   c. .....
   d. .....

TYPE 10. A spoken passage and a written summary with blanks. They must complete the blanks from the talk they have heard (Listening comprehension with reading comprehension!). However, the students could succesfully complete the written summary of the talk even if only little had been understood.

TYPE 11. The students hear a short talk or lecture and they are required to answer questions on it. Unless they are allowed to take notes, the test may put too heavy a load on the memory. First the students are given a sheet for note-taking.

Note-paper (It is given a few minutes before the talk)
You are going to hear a talk about ..... . After the talk you will be asked 25 questions about ..... . This sheet of paper is for any notes which you wish to take while you are listening to the talk. The questions you will be asked after the talk will be about the points listed below. A space has been left to enable you to write notes for each point.

(Questions are generally in the form of multiple choice / false-true items)