CHAPTER FIVE

Creating tasks

Introduction

In Chapter 4 I discussed the definition of a suitable listening construct. The next step in test construction is to operationalise that construct through a series of tasks to be carried out by the test-taker; not just any tasks, but tasks that are dependent on comprehension of the text, and whose successful completion requires the knowledge, skills and abilities we included in the construct definition. Then, from performance on these tasks, we can make inferences about how well test-takers have mastered the construct.

Most tests use a number of different task-types to operationalise the construct. The test specifications should lay out in detail what these various tasks should be, and each individual task may only operationalise part of the construct, but taken together the tasks need to represent the whole construct. To the extent that they do, the test will have construct validity.

In this chapter, I will discuss tasks from a variety of perspectives; it will be organised into four main sections:

i an overview of listening task characteristics;

ii the interaction between the task and the test-taker;

iii the use of comprehension questions;

iv evaluating and modifying tasks.

Chapter 6 will deal with issues related to constructing and delivering spoken texts; in this chapter discussion of texts will therefore be kept to a minimum.

Task characteristics

The framework of task characteristics given in Table 4.2 (p. 107) is intended to function as a checklist for comparing test tasks with target-language use tasks. It is a means of investigating task authenticity, as well as an aid to the development of new tasks (see Bachman and Palmer, 1996). In this section, I will discuss the components of the framework and consider how they apply to listening.

Characteristics of the setting

The circumstances in which a test is given can have a crucial effect on performance, and in the case of listening tests the most important characteristics of the setting must be those that affect the quality of the listening experience.

Physical characteristics

The physical situation is particularly important. It is necessary to ensure good acoustics and minimal background noise so test-takers can hear clearly and comfortably.

In some cases, the listening material will be presented by a live speaker which involves training speakers and making sure their presentations are as standardised as possible; but in most cases recorded sound will be provided, in which case the equipment is crucial. The recordings need to be clear, with a minimum of background hiss. The player needs to be of reasonably good quality, powerful enough to drive room speakers, and the playback heads need to be clean. Most importantly, the speakers need to be large enough to fill the whole listening space comfortably above the level of background noise, and without sound distortion. The same technical requirements apply if the recorded text is delivered via radio. In the case of video presentation or TV broadcast, it is important to provide listeners with an adequate view of the picture, with equal visual quality.

Language laboratories and computers usually deliver sound through individual headphones. This brings the sound right to the listener's ear, and can improve sound quality by reducing the effect of background noise. However, the headphones must be good enough for the job; poor headphones are just as bad as poor speakers.

Participants

Test administrators need to be efficient and know what they are doing. In the case of listening, this means they should understand any audio equipment being used. They should ensure that it is being used properly, and be prepared to modify sound levels so that everyone can hear comfortably. If headphones are being used, administrators need to be able to deal with defective ones.

Administrators must also ensure quiet, while the texts are being played: both by being quiet themselves and by keeping others quiet – especially those outside the room. This is not always easy, but high levels of noise can completely ruin a listening test administration.

Administrators who provide live texts to the listeners need to make sure they know what to do, and that they have practised and prepared themselves in the manner laid out in the instructions. They need to pay particular attention to any pauses inserted in the script.

Time of task

Generally, the time of the test administration will not be important, as long as the test-takers are not suffering fatigue. However, in some situations certain times may be preferable. Schools are noisy places when the students are all leaving, rush hours can lead to high levels of street noise, as can road works or street markets. Weekends, or evenings, might be much quieter, although administering tests at such times may not be possible.

Characteristics of the test rubric

The test rubric includes those characteristics of the test that provide structure to the test and the tasks. These have to be made explicit in the test, but are usually implicit in target-language use. However, the test rubric can provide a rationale for the activity and can function in an analogous way to the listening purpose in target-language use situations (Alderson, 2000). Test-developers can often increase the authenticity of the task, if they can structure the test rubric in such a way as to replicate the effects of a real-world listening purpose.

Instructions

As Brindley (1998) notes, it is important that test-takers are provided with clear, simple and explicit instructions on how to do the tasks. The penalty for misunderstanding one question is one item incorrect, but the penalty for misunderstanding one simple instruction could be many items incorrect – a penalty usually far out of proportion to the mistake made. Give instructions in the native language where possible, or if a second language is used, try to ensure that the language of the instructions is easier than the level the test aims to measure. Use clear examples, and if possible prepare the test-takers with sample items before they take the test.

In standardised tests it is likely that most test-takers will already know the instructions, and many of them will not pay attention to them again. If the instructions change, many test-takers may not notice, and the administration could go badly wrong. If reading or listening to the instructions is crucial, ensure that all test-takers are fully aware of this.

Test structure

The test specifications will specify the nature of the tasks, how many there are, and how they are organised and ordered. It is traditional to order test items with the easiest items coming first and harder items later. This ensures that lower-ability test-takers encounter all the easier items, which gives them the best chance of showing their true ability level. In the case of a **testlet**, that is a listening passage that has a number of items attached, it may not be possible to order the individual items according to difficulty, and the testlet must be treated as a whole unit.

Real-world listening always takes place within a context, and it is sometimes possible to turn the test itself into a context by organising the tasks around a theme. This creates a known situation that can be used to test context-embedded interpretations, sociolinguistic competence and inferences based on knowledge that has already been provided. For example, we could ask test-takers to imagine they were visiting England for a vacation, and then we could build a coherent set of tasks around that theme. Test-takers could listen to descriptions of hotel rooms; follow directions; listen to train timetables, theatre or concert times; listen to descriptions of places of interest or famous people; or they could listen to strangers talking about their work, their hobby or their views on social issues. The list is almost endless. After listening, test-takers could answer questions, fill in information grids or make decisions based on a specified context.

The theme does of course need to be relevant to the testing purpose. For example, in a test of business English, the theme could be a request from a colleague for information about a new product. Firstly, the listener could hear a message on a voice-mail asking for general information, followed by listening to a presentation about the new product, and then a talk on plans for product promotion. Such themes can easily be used to test integrated skills (Alderson, 2000). The listener could then be asked to read something about the new product, perhaps write a summary of the presentation, or record a spoken reply on the colleague's voice-mail.

There is much to recommend a thematic approach to test design, but the danger is that it will restrict the range of topics and text types, or advantage test-takers who have background knowledge relevant to the theme. When we develop a test, we want a sample of the competences and tasks in the construct definition. The greater and more varied the sample, the more likely it is to represent the construct fairly, and the smaller or narrower the sample the more likely it is to underrepresent the construct. Test-developers need to weigh the advantages of a thematic test design against the dangers of construct underrepresentation.

Time allotment

In most testing situations there is a predetermined time allotment, and the test administrators must keep track of time. In the case of listening, the timing of the test is usually determined by the sequence of texts and tasks. Often there will be one recording that includes all the listening texts, the instructions, questions, response pauses and so forth, and this recording will control the test; the administrator just presses the button and waits. It is easy to standardise administrations in this way.

Not all listening tests are controlled by one recording; sometimes there will be a number of recordings, and in other cases there may be a live presentation. In all cases, the test-developer will have to sequence the texts and tasks, and ensure there is enough time allowed for all the activities. For example, if we require test-takers to respond by making a simple mark on an answer sheet, five seconds' response time per item may be enough, but if we require them to write a sentence in response, then one minute may be necessary. Test-developers are well advised to practise the timing of activities, and trialling the whole procedure is advisable.

From the listeners' perspective, what this means is that they are not usually in control of their own speed of working. On a reading test, the test-taker can usually respond at a rate they feel comfortable with, and if they need more time for a difficult item, they can take it. This is not possible on many listening tests, and some test-takers may feel this is a disruption of their preferred way of working, and some may feel frustrated or even resentful as a result. However, some computer-based tests do allow the listener more freedom to decide when to move on to the next passage.

Scoring method

Scores are what give meaning to test performance, and our aim in test development is to ensure scores are meaningful in terms of the construct we have defined.

It is important that tasks are scorable; by that I mean that it must be possible to say, based on the listening passage, that certain responses are definitely correct and others are definitely incorrect. Given the nature of comprehension, the need to determine that one interpretation is correct and another is not correct can be problematic. When we insist on one particular understanding of a text, there must be no doubt that this was what the speaker intended. This is one reason why I have suggested in Chapter 4 that the default construct should test only those things for which there is clear and unequivocal linguistic evidence in the text. Although we would like to know whether listeners have understood the range of possible interpretations, any aspect of comprehension that is open to alternative interpretations is very difficult to assess.

When we require test-takers to construct responses, what they produce may be quite exhaustive or cursory, unequivocably correct or doubtful, indicate complete comprehension or only partial comprehension. It is important that the criteria for scoring these are clearly determined and consistently applied. For example, if a question asks why something happened, we would have to determine whether it is enough just to give the main point, or whether a full explanation is necessary. If there are two reasons in the text, do both need to be given in the response, is one required and the other optional, or is either good enough? One way of avoiding such problems is to use selected responses. The advantage is that they can constrain the responses available to the test-taker, who can be asked to choose between the correct response and alternatives that are definitely incorrect. This makes scoring much easier.

It is widely believed in educational testing that it is easier to write constructed response items but harder to score them, and conversely, it is harder to write selected response items but easier to score them. We write items once, but score them for each test-taker, therefore constructed response items will require less total work when testing volumes are small, and selected response items will require less total work when volumes are high.

Another important scoring issue is the extent to which the scoring criteria are made explicit to the test-taker. It is vital that test-takers know what constitutes a sufficient response. With open-ended responses they need to be told not only what is expected, but how much is expected, so that if they have understood the text well, they will be in no doubt about the adequacy of their reply. With multiple-choice tasks, there is less room for ambiguity, but instructions may need to ask for the best option, or the most appropriate response, rather than the correct one.

It is also important that test-takers know the relative value of each task in the test they are taking: how many points or what proportion of the total test each task is worth. This helps them to structure their time and efforts.

Characteristics of the input

The input into a listening task obviously consists of the listening text, but it will also include the instructions, questions and any materials required by the task. Much of the content of this section is dealt with in far more detail in Chapter 6.

Format

The input will be a sample of spoken language, and the length is important. Longer texts will tend to require discourse skills, whereas shorter texts will tend to focus more on localised grammatical characteristics. Listening to long texts can be tiring. Furthermore, when texts are even a little challenging to listeners, small difficulties can accrue in longer texts, until listeners lose the thread, get completely lost and just give up (Buck, 1990). It is certainly more difficult to maintain concentration if the content is boring or the presentation is bad. Interesting texts are much better. If the texts are challenging, it may be better to keep them shorter. If the specifications call for longer texts, it is often better to choose relatively easier texts. If the text is both longer and challenging, try to choose texts with an obvious structure, so that if listeners do lose the thread, they can get back into it relatively easily.

It is becoming increasingly common to present listening texts in video format. This is because visual information supplements the audio information, and this is a more realistic replication of real-world listening. It is important to make certain that the video complements the audio text, rather than conflicts with it, and that successful task completion still depends on processing the audio input. However, there can be problems: watching the video, reading the text, listening and writing responses all at the same time may be too complex, and many test-takers give up watching the video – which rather defeats the purpose (Alderson *et al.*, 1995).

The task will not only require a spoken text, but may also require written information, questions, pictures, diagrams, a grid to be filled in or a transcript of part of the text. In order to minimise construct-irrelevant variance, it is important that these are not a source of difficulty. Written language should be as simple and clear as possible – as should any pictures, diagrams and grids that are used. There are cultural differences in how things are conventionally represented in pictorial or diagrammatic form, and test-developers should ensure that test-takers understand clearly what is represented, and what they are supposed to do with it.

Language of input

The language of the input will be determined largely by the construct that has been defined. We need to ensure that the texts include the linguistic characteristics called for by the construct definition – aspects of grammatical, discourse, pragmatic and sociolinguistic knowledge. But that is not enough. It is also important that the test tasks engage the listener in *using* those competencies. For example, if the construct includes facility with a certain accent, it is relatively easy to check that the texts use that accent. Stress patterns and intonation are subtler, and it is not as easy to ensure that successful task completion requires understanding them. If the construct calls for discourse knowledge, just providing a longer text with comprehension questions may not actually engage the discourse characteristics; it might be possible to respond based on grammatical knowledge applied to short sections of the text.

Topical knowledge

Comprehension is an inferential process and the information that forms the basis for the inference is often outside the text. In some cases this may be part of the construct definition: for example, in a test of French for academic purposes we can assume that test-takers will know something about the French university system. In a test of German for business, we can assume that test-takers would recognise a German telephone number, or know how a fax machine works. This is construct-relevant knowledge. However, in other cases comprehension may require construct-irrelevant knowledge. There is ample theoretical evidence that background knowledge is important in listening comprehension, and there is research evidence to suggest that it does affect test performance (Chiang and Dunkel, 1992; Long, 1990; Schmidt-Rinehart, 1994). However, the relationship between background knowledge and test performance is complex and varies from one test to another, and probably from one topic to another (Jenson and Hansen, 1995).

There are three ways to ensure that test-takers have the background knowledge necessary: (a) use tasks that depend on knowledge that everyone has, (b) use tasks that depend on knowledge that no one has, or (c) use tasks that depend on knowledge that has been provided in the test.

Characteristics of the expected response

Responses can be provided in a wide variety of ways, but as discussed in Chapter 4, in all listening tests the response will be a potential source of construct-irrelevant variance.

Format

The format of the response can vary considerably. With selected responses it may amount to little more than a mark on a score sheet, or making a simple addition or alteration to a drawing or diagram. Constructed responses, however, vary greatly; these may require just writing one word in the case of a gap-filling test, or one or more sentences in the case of comprehension questions. A longer response could require writing a short summary of a text such as a lecture – this is not uncommon in research studies. It is also possible to have spoken responses, in sentence-repetition tasks for example, and listeners could be required to summarise what they have understood by speaking in their native language. Other response formats are drawing pictures or creating diagrams.

Language of expected response

In the case of constructed responses the language of the response is important. These could be given in either the target language or the test-takers' native language, and could be either written or spoken.

In the case of second-language responses, the main issue is whether the response will be evaluated for correctness, appropriacy and so forth, and whether those judgements will affect the score. In a listening test, it seems reasonable not to penalise mistakes if the response is intelligible and clear. After all, we want to know whether test-takers have understood a text, not whether they can produce correct language. Serious mistakes in the response will be penalised automatically, because when they are bad enough to interfere with the meaning, the response will be scored incorrect because it cannot be understood.

Task interactiveness

Looking at task characteristics is very useful, but it puts emphasis on the observable characteristics of tasks, whereas what we are really interested in is the knowledge and abilities that enable the test-takers to carry out that task. This brings us to the notion of **interactiveness**. This refers to 'ways in which the test-taker's areas of language knowledge, metacognitive strategies, topical knowledge, and affective schemata are engaged by the test task' (Bachman and Palmer, 1996:25). If the competence and knowledge engaged by the test task are those required by the construct definition, then the task has **interactional authenticity** in Bachman's (1991) terms, or minimal construct underrepresentation in Messick's (1989) terms. Put simply, it means the test is testing the construct it was intended to test. Interactiveness is important because it gets to the heart of construct validity.

We need to look at interactiveness from two perspectives: firstly, to see whether successful completion of the test task is dependent on comprehension of the text, and secondly, whether the knowledge, skills and abilities required to comprehend the passage represent the knowledge, skills and abilities in the construct definition.

Passage dependency

Passage dependency is the notion that successful completion of the task is dependent on comprehension of the text. It is central to comprehension testing. This is very basic and seems obvious, but a number of researchers have looked at multiple-choice reading tests and found that test-takers could usually score far better than chance even if they have not read the passage (Preston, 1964; Wendell *et al.*, 1967; Connor and Read 1978).

Test tasks may lack passage dependency for two reasons. Firstly, all test tasks provide some information to the test-taker, and this will usually give some clue to the content of the passage, which clever test-takers may use to help them respond. This is clearly the case with traditional multiple-choice items, but is also likely to happen with gap-filling tasks on written summaries, and can happen on information grids when part of the information is already given.

Secondly, tasks may lack passage dependency because test-takers might be able to use their background knowledge or intelligence to respond. Many things in life are predictable, and spoken texts are no different. Comprehension questions can often be answered by using common sense. For example in the case of a series of pictures that must be arranged in order based on a narrative, it may be possible to make sense of them without the narrative. In the case of multiple-choice options, a little thought will often suggest which options are more likely and which are less likely.

In Chapter 3 I noted that comprehension tasks must be designed in such a way that they provide clear evidence of comprehension. The only way to be sure the meaning has been processed is by using responses that require the semantic content to be expressed in a different form from the input. So for example, just pulling a phrase out of a text and repeating it does not provide sufficient evidence that the meaning has been processed, whereas matching the phrase to a picture, or matching it to a synonymous phrase does.

As test-developers, we should never underestimate the resourcefulness of test-takers in finding ways to compensate for a lack of comprehension. It is often difficult to ensure that our tasks are passage dependent, and that they provide clear evidence that the meaning has been processed.

Construct-relevant variance

We design tasks based on our construct definition, and different tasks will be appropriate for different constructs. For example, if we want to test a narrow view of listening – understanding short utterances on a literal level – we will look for a task that we believe requires this ability. Dictation or statement evaluation are possible choices.

Let us assume next that we want to assess a slightly broader construct: grammatical knowledge plus discourse knowledge, but not pragmatic or sociolinguistic knowledge. In this case we would have to find a longer text in which information was spread throughout the text, with plenty of references from one part of the text to another, and where the rhetorical structure was important. Alternatively, we could find a text with many oral discourse features, where the unplanned nature of the discourse resulted in repetitions, repairs or rephrasing of various parts. We then need to devise tasks that would engage the test-taker in understanding these aspects of the discourse. Neither the dictation nor the statement evaluation task would likely be suitable. Comprehension questions, statement completion tasks or gap-filling tasks could be used; the important issue is whether they engage the knowledge, skills and abilities of interest.

We could adopt an even wider view of listening. We may want to test grammatical, discourse and pragmatic competence, or the entire range of linguistic competence in the Chapter 4 framework. The procedure is the same: we first find texts whose comprehension requires these abilities, and then we devise tasks that engage them. Finally, in high-ability or first-language testing situations, we may want to go beyond the framework and test listening in the sense of understanding what the speaker felt about what she was saying. We would find a suitable text, or maybe a variety of short texts, in which the speaker's feelings are made evident by the tone of voice and the choice of words. We could then present listeners with a list of words describing emotions – angry, doubtful, fearful, embarrassed etc – and ask them to select the words that seemed most appropriate to the speaker's feelings. This is all easy in principle, but in practice it is difficult to determine exactly what sub-skills are required by particular tasks.

Construct-irrelevant variance

Whatever task we devise will require skills other than listening skills. In the case of the dictation, test-takers need to be able to write English; short-answer questions will require some reading and a little writing; and multiple-choice items usually require quite a lot of reading. As noted above, what is important is not that we get rid of these construct-irrelevant skills – that is not always possible – but to choose those that do not add any variance to item difficulty.

If we have a group of test-takers who are not literate, such as young children, tasks that require writing will not be appropriate. Test-takers who come from a first-language group which uses a different writing system, say Chinese students taking an English test, may be disadvantaged compared to test-takers who use the Roman alphabet in their own language, say French students. In the case of a test of Japanese as a second language, the same Chinese students, however, would likely be greatly advantaged compared to the French students, because both Chinese and Japanese use Chinese characters.

Multiple-choice items are another example. Many people believe that performance can be improved by the use of appropriate test-taking strategies, and there is evidence that practice does improve performance on multiple-choice items (Powers and Rock, 1999). Multiple-choice items are used widely in some countries, and virtually never in others. If we use them in multi-cultural situations, we may find some test-takers disadvantaged through lack of familiarity.

Task processing

Tasks are designed to engage particular aspects of language knowledge, but how do we know what knowledge, skills and abilities they require? The short answer is that we are never sure (Buck, 1991). Given the complexity of listening processes we can never be sure exactly which abilities particular tasks are engaging. Such labels as *vocabulary item* or *grammar test* rarely indicate more than the test-developers' intention. The reality will usually be far more complicated.

The necessary information

One useful way of focusing on what particular tasks are measuring is to identify the **necessary information**. This is the information in the text that the test-taker must understand in order to be sure the task has been done correctly. It is more than just the basis for a good guess; it is the information that would tell us whether the guess was correct. For some tasks, the necessary information may be one short statement, in others it may be a number of statements scattered throughout the passage, and in others it may be a whole section of text. Sometimes the necessary information is repeated in a passage, and sometimes it only appears once.

Once we have identified the necessary information, we can then assume that any knowledge, skills or abilities that are engaged in understanding this are likely to be important in task performance. Usually that is what the task will be assessing. Other abilities may also be involved, and there may be alternative ways an item might be processed, but the necessary information is usually the best and easiest way of developing hypotheses about what test tasks measure.

An example of how items are processed

I think it is always a good idea to examine illustrative items and attempt to understand how they may be functioning. The examples are taken from Buck (1991) and Buck (1994). Imagine a narrative about a woman who has been burgled many times, and she goes to a man called Dagbovie for help. The assumption up to this point is that Dagbovie is some sort of private investigator, but the story heads in a new direction when the narrator states, 'Dagbovie was a sort of animist or witch-doctor who had some special power to understand these strange events.' This was the necessary information that was targeted by two short-answer questions: Q1: *'What was Dagbovie?'*, and Q2: *'Why could Dagbovie understand strange events?'* The words *'animist'* and *'witch-doctor'* are quite low-frequency vocabulary, and these items were intended to test vocabulary knowledge. The expected response to Q1 was *'an animist'* or *'a witch-doctor'*, and to Q2 *'because he had some special power'*.

Six test-takers were selected and asked to discuss their listening and their test responses. The two highest ability test-takers responded correctly to both these items, and the four lowest ability test-takers responded incorrectly. This suggests these were both good items, because they discriminated well between listeners of different ability levels.

Now let us examine what happened. Firstly, regarding those four who got the item wrong, three of them heard the word *'witch-doctor'* and inferred that Dagbovie was a doctor. Furthermore, two of them misheard the word *'animist'* as *'analyst'*, presumably an inference based on the acoustic form of *'animist'* and perhaps due to some association with *doctor*. These two then inferred that Dagbovie was some sort of psychoanalyst. In other words, they did not understand the vocabulary being tested, and so they used their inferencing skills to compensate, and made very sensible inferences given what they knew – but they were wrong.

Now let us look at the two test-takers who responded correctly to these two questions. Can we assume they knew the vocabulary being tested? One responded that Dagbovie was *'someone with supernatural powers'*. In her discussion she explained that although she had correctly processed the words *'animist'* and *'witch-doctor'* she had no idea what they meant. She had answered correctly because she had understood the second part of this section that Dagbovie had some special power, and had made an inference that this was supernatural power. The other test-taker responded to Q1 with *'animist'*. The word was not in her vocabulary, but she inferred the meaning from the context.

Just like those who answered incorrectly, the two test-takers who answered correctly did not know the words *'animist'* or *'witch-doctor'* either, and they also used their inferencing skills to compensate for their lack of linguistic knowledge. Unlike the weaker listeners, however, their inferences were correct, presumably because they were processing the text efficiently, understanding it well, and they had more information on which to infer the meaning of the vocabulary items from the context.

To summarise, from the six test-takers we find three different reactions to Q1. The first group of two test-takers heard the isolated word *'doctor'* from *'witch-doctor'* and assumed that Dagbovie was a doctor. The second group of two test-takers misheard or inferred that Dagbovie was some sort of analyst, presumably based on hearing the word *'animist'*. The third group did not understand the vocabulary items but inferred what they meant from the content and their general knowledge.

Thus, if we consider the sub-skills involved in answering Q1, we must include at least the following:

- lexical knowledge
- the ability to isolate and recognise known words in an unprocessed stream of speech
- the ability to infer the form of words through phonological knowledge
- the ability to process clearly stated information
- the ability to ignore irrelevant information
- the ability to infer the meaning of words from context

- knowledge about the world in general
- and general knowledge about Africa.

If we now come to consider the sub-skills involved in answering the next question, Q2, we have to include all those which were involved in Q1, plus ideas about what constitutes a reasonable answer, clever guessing and more inferencing based on world knowledge, but also based on the inferences already made.

Looking at the relationship between answers to the two questions, it is easy to see how the comprehension and inferences made in one section, or in response to one question, could then in turn influence later sections. Thus, there is a cumulative effect, such that interpretations of later parts of the text, or test tasks coming later in the passage, can be influenced by a greater weight of personal, and perhaps idiosyncratic, inference than earlier items.

It is clear that we need to be very careful in labelling items in terms of the knowledge, skills and abilities we design them to measure. Processing language, as opposed to static linguistic knowledge, will always require a whole range of linguistic knowledge and processing skills, including phonology, vocabulary, syntax, discourse variables as well as inferencing ability. When the test-takers have difficulty comprehending, as will happen when we probe the limits of their knowledge in a testing situation, then processing becomes even more complex, and the result is a rich inter-linking of language skills with background knowledge, inference and speculation.

This does not mean we should give up designing tasks to address particular aspects of language competence. When creating tasks it helps to have some systematic scheme for item construction, and targeting particular components of language competence – such as knowledge of vocabulary, grammar, discourse etc – is one useful way of doing that. We are simply emphasising the things we think are important, but we must realise that the actual task processing will likely be quite different from what we intend. In the case of listening comprehension, task interactiveness is a complex and little understood process.

Targeting particular constructs

It would be very convenient to be able to offer a list of tasks suitable for each particular competence, or each definition of the listening construct, but as we have seen, task processing is often complex and unpredictable. Nevertheless, some tasks are more suitable for one purpose than another, and some suggestions can be made.

For example, although sound discrimination items are certainly not adequate tests of listening comprehension, they may still be perfectly good tests of knowledge of the sound system. There are plenty of situations in which language teachers will find their students having problems with the sounds of the target language, and they may want to devise tests that focus on these troublesome areas. Table 5.1 has a wide selection of tasks that are intended to do that.

Table 5.1. *Techniques for testing knowledge of the sound system*

Minimal pairs with decontextualised words

- Test-takers listen to two (or three) words and indicate whether they are the same or different.
- Test-takers listen to two (or three) words and indicate whether they rhyme.
- Test-takers listen to one word first, followed by a number of other words, and indicate which of these is the same as the first word.
- Test-takers listen to a number of words, some in the target language and some not, and indicate how many words are in the target language.
- Test-takers listen to two (or three) words and indicate which has the meaning expressed in a picture.

Minimal pairs with words in an utterance

- Test-takers listen to a sentence and choose which of two (or three) pictures indicates a word in the sentence.
- Test-takers listen to a statement, followed by two possible responses (which differ in only one word), and choose which response is appropriate.

Recognising grammatical structures

- Test-takers listen to a sentence and indicate whether the verb (or noun) was singular or plural.
- Test-takers listen to a sentence and indicate whether a particular word was masculine, feminine or neuter.
- Test-takers listen to a sentence and indicate whether the verb was in the past, present or future tense.

Recognising intonation patterns

- Test-takers listen to an utterance and choose one of three curved lines to indicate the intonation curve of the utterance.

Recognising stress

- Test-takers listen to an utterance, and then read three transcriptions of it in which capital letters indicate heavy stress; they choose the one which indicates the stress of the utterance they listen to.

Adapted from Valette (1977)

Another definition of the listening construct is understanding explicitly stated information, or understanding the passage on a literal semantic level. It is relatively easy to assess this, and Table 5.2 lists a number of suitable tasks.

It is much more difficult to find tasks which are primarily suitable for testing a broader listening construct or going beyond literal meanings. Probably the most common way of testing these skills is by asking comprehension questions. These are discussed in detail in the next section. However some tasks do suggest themselves, and these are given in Table 5.3.

The tasks in Table 5.3 do not automatically test listening in the wider sense, but they have the potential to do so. It is up to the test-developer to use them to test such skills as understanding gist, inferences, implied meanings, contextual implications and so forth.

Comprehension questions

Probably the most common tasks used to assess listening are comprehension questions. The idea is simple: a text is presented to test-takers, who are then asked to answer questions designed to measure how well they have understood the content. This is a common procedure; comprehension questions can be used with a variety of text-types, and they can be used to test a wide range of knowledge, skills and abilities. Furthermore, they appear easy to devise, although as we shall see, there can be problems.

Some general points

Before looking at particular question types, it is useful to consider some general points. The first issue concerns the type of information that should be targeted. Shohamy and Inbar (1991) looked at three types of questions: global questions, which required test-takers to synthesise information or draw conclusions, local questions, which required test-takers to locate details or understand individual words, and trivial questions, which required test-takers to understand precise but irrelevant details not related to the main topic. They found that the global questions were harder than the local questions, but that the trivial questions behaved in an unpredictable manner

Table 5.2. *Tasks for testing understanding of literal meanings*

Body movement tasks

- Test-takers are given commands to move various body parts (raise your right hand, go to the door etc.).
- Simon says: test-takers are given commands to move various body parts (e.g. raise your right hand, go to the door etc.) and only do so if the command is preceded by 'Simon says'.
- Test-takers are told to draw a certain object (e.g. draw a rose, draw a circle if you are a girl etc.).

Retention tasks

- Test-takers listen to two utterances and indicate whether they were the same or different.
- Test-takers listen to an utterance and write it down.

Picture tasks

- Test-takers listen to a statement, and then look at a number of pictures and indicate which represents the statement.
- Test-takers look at a picture and listen to a number of statements, and then choose which is relevant to the picture.
- Test-takers are given a pile of about ten pictures and listen to a series of utterances; they have to choose a picture of the object mentioned in each utterance.
- Test-takers are shown a complex picture with many things happening in it, they hear a series of statements about the picture and indicate whether these are true or false.
- Test-takers look at four pictures labelled A, B, C and D, and listen to a series of utterances (e.g. it is raining, the man is wearing a suit) and indicate which picture each utterance applies to.
- Test-takers look at five small pictures, listen to four short utterances, and select the picture being described in each utterance.
- Test-takers see a series of simple diagrams (e.g. lines, squares, rectangles, circles etc. arranged in a variety of patterns), listen to statements describing these, and indicate which diagram is being described.

Conversation tasks

- Test-takers listen to a statement followed by a response, and indicate A if the response was appropriate, and B if it was not.
- Test-takers listen to a question followed by three possible responses, and indicate which was the most appropriate.
- Test-takers listen to a statement followed by three possible continuations, and select the one that would continue the conversation best.
- Test-takers listen to a short dialogue (usually persons A, B and A again), and then listen to a short question about it, to which they choose one of three (or four) options (usually written in a booklet).

Self-evident comprehension tasks

- Test-takers are given a series of statements and asked to indicate whether they are true or false (e.g. the snow is white, Paris is in Spain).
- Test-takers listen to statements made about some sort of visual (or object), and indicate whether they are true or false.
- Test-takers are given arithmetic calculations and indicate whether they are right or wrong (e.g. three plus three is seven).

Adapted from Valette (1977) and Heaton (1990)

Table 5.3. *Tasks for going beyond local literal meanings*

Understanding gist

- Test-takers listen to recordings from the radio, and indicate what type of programme they are listening to (e.g. news, sports, weather, fashion etc.).
- Test-takers listen to a description of a person, place or thing they are all familiar with, and write down its name, or what it is (perhaps in their native language).

General passage comprehension

- Test-takers listen to a monologue or dialogue, and then answer a number of multiple-choice questions (either spoken or written questions).
- Test-takers listen to a monologue or dialogue, and then answer a number of short-answer comprehension questions on it.
- Test-takers listen to a short talk or lecture, read a number of statements and select which are correct given the talk.
- Test-takers listen to a short talk or lecture and answer comprehension questions on it.
- Test-takers listen to a short talk or lecture, and then read a short summary of it in which some words or phrases have been deleted and replaced with blanks; they fill in the blanks with content appropriate to the talk.

Information transfer tasks

- Test-takers are given a map and a starting point, they follow directions on the map, and then indicate the destination (the instructions can be extremely complex and as an example Heaton (1990) recommends listening to a description of a robbery and then following the police chase, periodically indicating where the robbers are).
- Test-takers listen to an announcement of some information (e.g. a timetable or result of a competition) and fill in the information in a grid.
- Test-takers are given a goal (e.g. arrive in Paris before midnight, solve a problem), they listen to an announcement or explanation, find the information necessary to complete their goal, and then do the task.
- Test-takers listen to a person on the telephone and take a message according to the speaker's instructions.
- Test-takers look at a picture (e.g. a street scene, a room) and draw in objects or other details according to instructions (draw a table and two chairs in front of the café; draw a vase of flowers on top of the table); this can be expanded into a narrative (or dialogue).
- Test-takers see a series of pictures and listen to a talk or discussion and then classify the pictures based on that (e.g. a set of children's drawings representing different developmental stages, which need to be put in order based on a talk about child development).

Adapted from Valette (1977) and Heaton (1990) and 'served no meaningful purpose in an evaluation tool' (Shohamy and Inbar, 1991:37). This research suggests that questions need to focus on the key information in the text, not irrelevant detail.

Another important issue is whether the questions should be provided before listening or after. In virtually all real-world listening situations, listeners know why they are listening and what information they need to get from the text. If we ask test-takers to listen to a text with no specified purpose, and after listening give them questions that require understanding anything and everything, we are asking them to do something very unnatural. Hence, most scholars recommend giving test-takers the questions before listening (Berne, 1995; Brindley, 1998; Buck, 1990). However, research on the effects of question preview is rather mixed. Buck (1990) made two versions of a test, one with item preview and one without preview, and gave them to two groups of test-takers. Although test-takers clearly felt that question preview aided comprehension, results showed that it did not make a significant difference to item difficulty. However, question preview did make a considerable difference to item discrimination, suggesting that preview does affect item processing, but in ways not yet understood. Berne (1995) found that question preview led to significant improvements in test performance compared to the effects of other pre-listening activities, whereas Sherman (1997) found that preview did not significantly improve performance. Wu (1998) found that preview of multiple-choice questions facilitated processing for the more advanced listeners, but not for the less advanced.

It is very difficult to make firm recommendations about question preview, but it does seem to have a positive psychological value for test-takers, and conversely the uncertainty of not knowing why they are listening has a negative psychological effect (Buck 1990; Sherman, 1997). Another problem is that a list of questions often makes little sense out of context, and it is not until listeners hear the passage that the questions start to seem relevant (Sherman, 1997). Question preview may not motivate listening as much as test-developers hope. Sherman (1997) found that the most powerful improvement in performance came from showing the questions after listeners had heard the passage once, but before they heard it a second time.

There is some evidence to suggest that if a listening passage is intrinsically interesting, there is no need to motivate the listening with questions – listeners will simply listen out of interest (Buck, 1990). If the questions then focus on the main topic, or the central thread of the discourse, good listeners will have the information necessary to respond, in which case question preview may not be necessary. If, however, the test-developer doubts whether the passage is intrinsically interesting, or doubts whether the questions focus on the main points of interest, then providing question preview seems the most prudent course of action.

There are some conventions about the use of comprehension questions. It seems to be widely accepted that questions will ask for information in the same order in which it occurs in the passage (Thompson, 1995), and also that once an item of information has been used in one question, it will not be used again in another (Buck, 1990). Test-developers who break these conventions may confuse test-takers, which could lead to unreliable performance.

In many testing situations, especially testing academic listening, test-takers want to take notes while listening. Whether this is allowed or not will depend on the definition of the listening construct and target-language use situation. Hale and Courtney (1994) found that allowing students to take notes during listening did not significantly improve their performance on the TOEFL listening test, and urging them to take notes significantly impaired their performance.

Finally, there is the issue of question spacing. It is important to leave enough time for test-takers to respond to each question, before going on to the next. This is especially problematic when test-takers are asked to complete a task at the same time they are listening – especially when filling in information grids, but this also happens with short-answer questions or multiple-choice questions. As Brindley (1998) notes, it is very important that questions are spaced out so that test-takers will not be responding to one question at the same time as they need to be listening to the information necessary for the next one.

Short-answer questions

Comprehension questions are relatively simple to write if we allow test-takers to construct the responses themselves. If we design the questions so that the answers will be short, scoring should be reasonably fast. Such questions have much to recommend them, and are particularly suitable for testing the understanding of clearly stated information. We can illustrate the process by attaching questions to a passage from Chapter 2 where the young man talks about a tennis tournament.

Example 5.1

Test-takers hear:

> *and i got there, and it turned out i was only the second person signed up for the entire tournament. and so there_ the people who ran it were like 'do you even want to bother playing the tournament?' and i was like_ 'let's it a great opportunity because if i win then i've won a tournament only with one match. so i can beat one guy, and even if i don't, like who cares.' so_ and_ cuz you get points for when you finish. so a runner-up gets the points anyway. so i was kind of hoping that the other guy i was gonna playing against was gonna dropout, and just give me the trophy. but he didn't. so anyway we went out and played, and we didn't even start_ we started like way after_ because i got there for a seven o'clock match. we didn't start playing until like ten at night.*

Test-takers read:

Q1. How many people turned up to play?
Q2. Did the speaker want to play?
Q3. Why was he not very worried about losing?
Q4. How late were they in starting the match?

Sometimes short-answer questions can be presented in a sentence-completion format. Example 5.2 gives sample items from the December 1998 listening paper of the Cambridge First Certificate (reviewed in Chapter 8).

Example 5.2

9. Emily's childhood and background were not very ________________
10. Unlike her brothers, she was educated ____________________________
11. With her tutor, she travelled around ______________________________
12. She then produced ____

These sentence-completion items could just as easily be presented as short-answer comprehension questions. However, the fact that these items are easy to produce, does not mean that it is easy to produce good ones. Like all test items, there are pitfalls, and items need to be pre-tested and bad ones need to be identified and either deleted or re-written.

Marking

Allowing test-takers the freedom to construct their own responses means that we need to make decisions about which responses will be marked correct and which will not. This is the most difficult aspect of using open-ended questions. In the case of items that ask for a simple fact, scoring is not usually so difficult. For example if the passage says, *'the train left at ten o'clock'*, and the question asks, *'What time did the train leave?'*, there is little chance of ambiguity.

The way to make marking easy is to write items that have unambiguous responses. This is often quite possible, and many texts will allow plenty of such items. However, the result may be questions that focus on superficial understanding of clearly stated information, and as test-developers, we are often trying to assess comprehension at a subtler level. As we probe deeper, however, it becomes more difficult to determine what should constitute a correct response. There are two potential problems: firstly, determining what constitutes a reasonable interpretation of the text, and secondly, what constitutes a sufficient response to the question.

Evaluating responses

Imagine a test about a burglar who *'was stealing something very small'*, and then a question asks, *'What was taken?'* The obvious answer would seem to be *'something small'*. So how should we score the answer *'something trivial'* given by a person who assumed that something small meant something of no value? Is this a correct response or not? What about a response *'nothing of value'*? When we know that some small things could actually be quite valuable, it is not easy to determine which should be accepted as correct and which not.

There is also the problem of deciding what is sufficient information to constitute a full response. Imagine that the same text continues by saying that the woman tried many things to stop the burglaries – putting bars over the windows and hiring a guard. If the question asks, *'What did she do about her problem?'*, what is the correct response? Is it enough to say she tried to stop the burglaries, or do we insist that the test-takers say how she tried to stop them? After all, we could guess that she would try to stop them, without listening to the text. And how will the test-takers know their response is good enough? These are difficult decisions that should only be made in the context of the testing situation.

A strategy for determining correct responses

The fact is that comprehension is not a simple dichotomy, and **dichotomous scoring** (i.e. scoring each item as being either correct or incorrect) of free-response questions turns a complex continuum into a simple dichotomy. The scorer has to make a whole series of decisions about which responses are acceptable, and which are not.

How can we decide? With sufficient resources the test can be administered to a group of competent listeners and their responses used as a basis for judging the acceptability of responses; but many of us will not have sufficient resources to trial our items in this way. In that case, when scoring the items, keep a record of the responses: note which are definitely right, which are definitely wrong, and which are borderline. Keep a note of the papers which have the borderline decisions. As you continue scoring you will come to understand better how the test-takers are interacting with the task and what the question is measuring, and as a result you will probably develop a better idea of exactly which of the borderline cases should be accepted as correct and which should not. Then go back and re-mark the borderline responses. Slowly a reasonably consistent marking scheme should emerge, even though there may still be some inconsistencies. This is not an ideal way of scoring, but in the real world of limited resources, we cannot always pre-test everything as we would like.

Shortage of time

In any test, listeners need to know how much time they have to respond, and this is particularly important with short-answer questions, because there is usually only time for short answers. If test-takers give answers which are too long, they may spend too much time on one question, and could miss the next question, or even the next listening passage. A useful strategy is to restrict the response to no more than three words, for example, although this is not always practicable.

Sometimes test-takers take a long time over a question because they are thinking about the response, but generally, if test-takers have understood the passage, thinking time will be short compared to the time taken to write the response. So when test-takers are spending extra time thinking of a response, it is usually because they have not understood the text well enough to respond. If individual test-takers want extra time to think about their responses, assume their comprehension is lacking, and find some way to force them to move on.

Longer responses

Sometimes we may want to ask comprehension questions that require longer responses. For example, if we asked why something happened, we might get two or three sentences in response. When responses are longer and more complex, we will have to develop some sort of rating scale to evaluate their suitability.

Developing a rating scale involves deciding what responses test-takers are likely to produce, how these should be evaluated relative to each other, and how many marks should be awarded for each. Having developed a scale, the job has only just begun. It is then necessary to try it out on a sample of tasks, preferably with a number of raters. Then all raters must practise using the scale until they can all apply it consistently. Developing a reasonable rating scale is not usually difficult, but it does require time and effort to develop it properly (McNamara, 1996).

Multiple-choice questions

Selected responses can be of many types, but the most common is the multiple-choice item with three, four or even five options. Constructing these items is a high-level professional skill, which takes considerable time and training to do well. All items ought to be pre-tested before being used in any high-stakes assessment, but this is particularly the case with multiple-choice items. They are complex and unpredictable, and after pre-testing expect to throw away, or seriously modify, a large proportion of your multiple-choice items, perhaps even all of them. Most people are very disappointed with their first attempts to write multiple-choice items.

Multiple-choice items are difficult to write because of their complexity. Many scholars believe that they have a strong method effect (Brindley, 1998) and that they make considerable processing demands on the test-taker (Hanson and Jensen, 1994). They can force test-takers to re-adjust their interpretation if it does not agree with the options (Nissan, DeVincenzi and Tang, 1996). Wu (1998) found that they favoured more advanced listeners, and that misinterpretations of the options led to test-takers selecting the wrong options, and conversely, test-takers selecting the correct options for the wrong reasons.

If the test-takers are all from the same L1 background, giving the questions in the first language works very well, and is probably the best way of ensuring that listening test scores are not contaminated with other skills. However, in many cases test-takers will be from different first-language backgrounds, and it will be necessary to give questions in the L2. The most common way is to provide written questions, and so multiple-choice listening tests are tests of reading as much as listening – which often may not be a bad thing.

There are two basic formats for presenting the questions. In Example 5.3 below the stem is in the form of a question, followed by alternative responses.

Example 5.3

Test-takers read:

When did she get back home from her sister's house?

(a) *Just before six o'clock.*
(b) *Late in the afternoon.*
(c) *About tea time.*
(d) *At exactly half-past nine.*

An alternative is the sentence-completion format, where the stem is an incomplete sentence, and the options are alternative completions. Example 5.4 gives the same item in this format.

Example 5.4

She got back home from her sister's house

(a) *just before six o'clock.*
(b) *late in the afternoon.*
(c) *about tea time.*
(d) *at exactly half-past nine.*

Example 5.5 is a typical multiple-choice comprehension question, taken from a TOEIC sample test (TOEIC MT-93, 1993). Test-takers hear a short dialogue, and then respond to a comprehension question based on the dialogue.

Example 5.5

Test-takers hear:

Female:	*is Carlos really quitting?*
Male:	*yes he's tired of the restaurant business. he wants to try something entirely different.*
Female:	*that leaves his sister Rosa to run it alone.*

They read:

Why is Carlos changing his job?

(a) *He is lonely.*
(b) *He is moving to another city.*
(c) *He has lost interest.*
(d) *He is ill.*

This item requires understanding a number of short utterances on a literal semantic level. The test-taker must equate the meaning of *'he has lost interest'* with *'he's tired of the restaurant business. He wants to try something entirely different'*. The two expressions mean the same thing, but they are not simple synonyms. Example 5.6 is from the Test of English as a Foreign Language website (http://www.toefl.org/lc-pq.html).

Example 5.6

Test-takers hear:

Male:	*do you mind if i turn the television off?*
Female:	*well i'm in the middle of watching a program.*
Narrator:	*what does the woman imply?*

They read:

(a) *The man should watch the program too.*
(b) *The man should leave the television on.*
(c) *The program will be over soon.*
(d) *She'll watch television later.*

This example is different for two reasons. Firstly, the question is presented aurally (note that it helps if the narrator's voice is noticeably different from the speakers'). Secondly, the question is given after the text has been heard, so test-takers will not know what to listen for before they hear the text (when the questions are written in the test book, test-takers can often take a quick peek before listening).

Another interesting point about Example 5.6 is that the woman does not explicitly refuse the man's request, so the listener must go beyond understanding the literal words and make a pragmatic inference in order to understand that the woman is actually refusing permission to switch off the television. Although this is an easy inference to make, the test-taker has to understand the inferred meaning by taking into account the communicative situation.

In the next example, also from the TOEIC sample test (TOEIC MT-93, 1993), test-takers hear a short talk, and then answer comprehension questions.

Example 5.7

Test-takers hear:

> *welcome to our semiannual sales meeting everyone. after lunch and a brief business meeting, a team from our research and development department will join us and demonstrate our newest products. each of you will have chance to try samples from our new line and ask questions of the team. now please help yourselves to the delicious buffet that has been set up in the adjoining dining room.*

They read:

Q1: *Who is attending the meeting?*

(a) *Sales personnel.*

(b) *Food service staff.*

(c) *Bank executives.*

(d) *Factory workers.*

Q2: *What will the people do first?*

(a) *Try out some new products.*

(b) *Eat a meal.*

(c) *Visit the research department.*

(d) *Discuss salaries.*

These two items are clearly testing comprehension. The questions are printed in the test book, and so test-takers may have time to scan the items first, which means they should be able to listen for specific information, rather than just generally trying to understand everything. Q1 *'Who is attending the meeting?'* requires understanding a very short piece of explicitly stated information, *'welcome to our semiannual sales meeting'*, or perhaps it is enough just to catch the expression *'sales meeting'*. Q2 *'What will the people do first?'* requires test-takers to understand that the expression *'now please help yourselves to the delicious buffet that has been set up in the adjoining dining room'* is an instruction to go and eat. Again not a difficult inference, but a pragmatic inference, nevertheless.

It is outside the scope of this book to go into details of how to write good multiple-choice items (see Popham, 1989), but a few words of advice might help. Make sure the question is about something on which there could be plausible alternatives. It is often the case that there is one obvious distractor, which is definitely wrong, but which seems likely to attract students who do not understand. It is much harder to find the second and third distractors. An option that no one chooses is worthless. When looking for alternative options, think of other information similar to that in the correct option, or try words that sound similar. Something that would appear plausible to a person who did not hear the text will often work.

Make sure that the form of the question does not give away the correct option. Check that all the options fit the stem, and that the correct one is not longer, nor different in any way. Make sure that neither the stem nor the options are ambiguous, and check that none of the incorrect options could be considered correct. Take care that the answer cannot be provided from background knowledge or basic reasoning.

Although they are complex and difficult to make, multiple-choice items can be used to test a variety of listening sub-skills: from understanding at the most explicit literal level, through combining information from different parts of the text, making pragmatic inferences, understanding implicit meanings, to summarising and synthesising extensive sections of test.

True/false questions

There is another type of selected option question that is very popular, the true/false format. This is very simple: after presenting a text, one or more statements are given, and test-takers have to decide whether each statement is true or false. Example 5.8 illustrates the format with a series of true/false questions from the December 1998 CCSE.

Example 5.8

After listening to the statements, test-takers mark whether they are true or false.

	True	**False**
1 Barry never sleeps more than 5 hours.	□	□
2 He wishes he could sleep longer.	□	□
3 He thinks vegetarians may need more sleep.	□	□
4 His fitness is not affected by his sleeping pattern.	□	□

There is some disagreement about the utility of this question type (Brindley, 1998). Burger and Doherty (1992) claim that they are not suitable for testing listening because listeners normally focus on what is said, not on what is not said, and as there is no text for them to refer back to, listeners have no means of checking false statements.

The other major problem with true/false questions is that test-takers can get half the questions correct by random guessing. This is also a problem with three- or four-option multiple-choice questions. There are a number of ways of dealing with guessing. Firstly, it is possible to reduce the effects of random guessing by including more items on the test – the more items there are, the less effect each individual guess has on the total score. A second problem arises when some test-takers are guessing and others are not. It is important to tell test-takers that they should guess, so that they are all pursuing the same strategy.

There is another advantage to having test-takers guess. Many guesses are not purely random guesses at all. Even when they do not know the correct answer, test-takers often have some comprehension of the point being tested, and so they are inclined to favour one response over another, perhaps without even knowing why. In such cases, the guess is based on some degree of partial comprehension which is relevant information about the construct being measured.

Inference questions

During the course of the previous chapters, I have continually claimed that inference is at the core of language processing. Thus, it is important to assess inferencing ability, and many scholars recommend questions that go beyond literal meanings (Boyle and Suen, 1994; Burger and Doherty, 1992; Thompson, 1995; Weir, 1993). Inference questions can, however, be difficult to make because the answers are not explicitly stated in the text.

It is important to distinguish between two types of inferences in test tasks. The first is inferences about what the speaker means. These are usually construct-relevant inferences. The second are inferences about what the test-developer expects, and what the best test-taking strategy is. These are usually construct-irrelevant inferences, and we should try to keep them out of our tests.

The main problem with inference items is that it is often very difficult to say with certainty that any inference is completely wrong. After all, we are free to infer what we want. There are other problems. It must be possible to get the item wrong – items which ask for personal reactions or predictions about future events clearly cannot be marked wrong, for example. Also, questions must be passage dependent, in the sense that they require something which can only be inferred by understanding the passage. It is often possible to use common sense or general knowledge to infer a reasonable reply. Furthermore, inferences are based on background knowledge, and it is important to make sure that the knowledge is shared by all test-takers.

Many of these problems can be avoided by using multiple-choice items. Even if the correct response is not the best possible inference, as long as the other options are clearly less preferable, it should be obvious which is the correct response. Make sure to ask test-takers to choose the *best* option, rather than asking them to choose the *correct* option.

The sort of information that can usually be addressed by inference questions is:

- asking for the main idea, or gist, of the text, or a section of the text;
- asking about anything which is not clearly stated, but that is clearly and deliberately indicated by the speaker, using choice of words or tone of voice – the connotations of words is a particularly rich source of inferences;
- asking about any pragmatic implication, or logical entailment, that follows on from what the speaker said;
- asking the meaning of indirect speech acts.

Inferencing is involved at all levels of language processing, even explicitly stated information. Once we attempt to test pragmatic or sociolinguistic knowledge, which involves interpreting meaning in terms of a wider communicative context, then inferencing becomes even more important. Good listening tests will generally include inference items. It takes time and trouble to make them, but they are too important to exclude.

Evaluating and modifying test-tasks

Creating test tasks is a complex skill, and there are many ways they can go wrong: for example, there may be no correct answer, there may be two answers, or it might be possible to answer correctly without understanding the passage. The best way to avoid such problems is to give the tasks to a small number of potential test-takers, or colleagues, ask them to complete the task, and then solicit their feedback. This will usually reveal any obvious problems.

This will not provide information about the difficulty of the task, nor whether it is measuring the right construct. In order to do that, it is necessary to pre-test the items on a sample of test-takers similar to the target population, and then subject the results to item analysis. This pre-testing will show whether the tasks are of the right difficulty level, and whether they discriminate well between test-takers who have different abilities on the construct. After pre-testing, it is usually worthwhile trying to modify problematic tasks so they can be used. In the case of poor discrimination this can be problematic – in some cases it is very clear what is wrong, whereas in others it may not be clear at all.

In the case of difficulty, it is often possible to modify tasks to make them easier or harder. This can be done by modifying the text, or by modifying the task. In the following sections, variables that affect text difficulty and variables that affect task difficulty are examined, in order that task difficulty can be better evaluated, or in order to make test tasks more suitable for the target test-takers.

Text characteristics that affect difficulty

The following list is put together from a number of different sources, including Anderson and Lynch (1988), Brown (1995b), Freedle and Kostin (1996, 1999), Buck *et al.* (1997), Buck and Tatsuoka (1998).

Linguistic characteristics

- Texts with slower speech rates tend to be easier than texts with faster speech rates.
- Texts with longer pauses between idea units tend to be easier than texts with shorter pauses between idea units, or no pauses at all.
- Texts with more familiar pronunciation tend to be easier than texts with less familiar pronunciation.

- Texts with natural intonation patterns tend to be easier than texts with unnatural or unusual intonation patterns.
- Texts with more high-frequency vocabulary (i.e. common words) tend to be easier than texts with more low-frequency vocabulary.
- Texts with less complex grammar tend to be easier than texts with more complex grammar.
- Texts with idea units or clauses strung together tend to be easier than texts with ideas units or clauses embedded within other clauses.
- Texts with simple pronoun referencing tend to be easier than texts with more complex pronoun referencing.

Explicitness

- Texts in which the ideas are explicitly stated tend to be easier than texts with less explicit ideas.
- Texts with more redundancy tend to be easier than texts with less redundancy (but not if the listeners fail to realise that the information is redundant).

Organisation

- Texts which have events described in a linear or temporal order tend to be easier than texts which have non-linear structure.
- Texts which have main points stated clearly before examples tend to be easier than texts with illustrative examples coming before the point being made.

Content

- Texts with topics more familiar to the listener tend to be easier than texts with less familiar topics.
- Texts with fewer things or people to be distinguished tend to be easier than texts with more things to be distinguished.
- Texts in which the important protagonists or objects are more easily distinguished tend to be easier than texts where they are harder to distinguish.
- Texts where relationships between the elements are fixed tend to be easier than texts where relationships are changing.
- Texts with concrete content tend to be easier than texts with abstract content.

Context

- Texts with visual or other support which supplements the content tend to be easier than texts with visual or other information that conflicts with the content.

Task characteristics that affect difficulty

Probably the simplest way to make tasks easier is to allow listeners to hear the text twice (Berne, 1995). However, playing the text a second time may significantly change the nature of the listening construct, and such a decision should probably be made as part of the process of construct definition, rather than as part of task revision. Tasks can be made even easier by giving the questions or the task in the interval between the two hearings (Sherman, 1997). Below are some general guidelines taken from a variety of sources.

- Tasks that require processing less information tend to be easier than tasks which require processing more information.
- Tasks that require processing information from just one location in the text tend to be easier than tasks which require integrating information scattered throughout the text.
- Tasks that require recalling exact content tend to be easier than tasks which require extracting the gist, or making a summary.
- Tasks that require simply selecting information tend to be easier than tasks which require separating fact from opinion.
- Tasks that require information that is relevant to the main theme tend to be easier than tasks which ask for irrelevant detail.
- Tasks that require immediate responses tend to be easier than tasks which require a delayed response.

Research on variables that affect difficulty

We can see many of these characteristics in action in a study by Freedle and Kostin (1996, 1999). They examined 337 TOEFL multiple-choice listening items, which asked comprehension questions on short monologues, in order to examine which item characteristics were related to difficulty.

They found that questions which required understanding explicit statements were easier when:

- the necessary information was repeated;
- the correct option contained more words used in the necessary information than the incorrect options did;
- the correct option contained more words used in the text than the incorrect options did;
- the incorrect options contained more complex grammar than the correct options did;
- there were relatively more topic shifts in the passage;
- the text was on a non-academic topic;

and they were harder when:

- the passage had relatively more text coming before the necessary information;
- the question had relatively more referentials (such as pronouns).

In the case of the items which required the test-taker to make inferences, they found items were easier when:

- the necessary information came at the beginning of the passage;
- the correct option contained more words used in the text than the incorrect options did;
- the topic was arts, or social science;

and they were harder when:

- the necessary information was not repeated;
- the necessary information came in the middle of the text;
- the rhetorical structure of the passage involved a comparison.

Finally, items which asked for the identification of the main idea were easier when:

- the topic was a non-academic topic;
- the correct option contained more words from the text than the incorrect options did;

and they were harder when:

- the rhetorical structure of the passage was problem solution.

A number of things are clear from this study: the topic affected difficulty, as did the rhetorical structure, but the two most important determinants of difficulty were the location of the necessary information, and the degree of lexical overlap. When the necessary information came near the beginning of the text, or when it was repeated, the item tended to be easier. **Lexical overlap** is when words used in the passage are found in the question or in the options. Lexical overlap between the correct option and the text, especially the necessary information, is the best predictor of easy items. Similarly, lexical overlap between the text and the incorrect options is the best predictor of difficult items. Presumably, this is because test-takers tend to select options which contain words they recognise from the passage.

Conclusion

Having defined a listening construct, the next step is to operationalise that in terms of tasks. All tasks have their particular strengths and weaknesses, and tend to engage different skills: some assess one set of skills and some another; some advantage one group of test-takers, and some another. By using a variety of different task types, the test is far more likely to provide a balanced assessment, and it will usually be a fairer test (Brindley, 1998).

In this chapter I have discussed tasks from a variety of perspectives; firstly, in terms of the Bachman and Palmer (1996) schedule of task characteristics. Then I argued that the most important characteristic of tasks is the knowledge, skills and abilities they engage in the test-taker, and I discussed at length various aspects of task interactiveness. The point was made that although we would like to create tasks to address particular sub-skills of listening, task performance is essentially unpredictable. Nevertheless, certain tasks were examined as being suitable for testing certain aspects of listening. I also discussed comprehension questions in detail, and considered how to make texts and tasks easier or harder.

In the next chapter I will discuss texts; how to select them, or create them, and how to present them to the test-taker.