3 Item writing and moderation

In this chapter we shall be discussing what is involved in writing good test items. We shall describe some of the pitfalls to avoid and the procedures to follow that will help ensure that many obvious mistakes are caught before the test is pretested. We shall attempt to answer the following questions: What makes a good item writer – are they born or can they be trained? Where do you start when writing an item? What methods are most suitable for testing particular abilities? When people disagree about the quality of a test item, how can we resolve the disagreement? What principles and guidelines should we follow when writing test items? What is the role of the moderating committee, and how do such committees best work?

3.1 Qualifications for item writing

The purpose and content of the test will to some extent determine who will make the best item writers. It is helpful if those who write the items have recent experience teaching students who are similar to those taking the test, as the teachers' experience will provide insights into what such students find easy and difficult, what interests them, their cultural background, and so on. For example, if the test is one of Writing for Academic Purposes, then a person with experience of doing academic writing, teaching academic writing, and assessing students who submit pieces of academic writing is likely to produce a better test of academic writing than somebody without such experience. For achievement tests, it is clearly important that those who write the test know what it is reasonable to expect students to have covered at that particular stage in learning and also how far students have actually progressed through the curriculum. Thus it is likely that good item writers will be experienced teachers of similar students or relevant subject areas, who have the necessary formal professional qualifications expected of teachers in the particular context where the test is being developed.

However, such people will not necessarily be good item writers.

Having relevant experience does not guarantee either the insights into the nature of the task that are necessary in order to write good tasks, or the creativity and imagination needed to write good items. Creativity, sensitivity, insight, imagination: all these are qualities of item writers that are difficult to define and difficult to identify in prospective item writers, but very obviously missing in poor item writers.

Some tests are written by professional item writers, who either work full-time for a testing institution, or who work as freelance test writers for a range of institutions. Such writers should ideally combine the experience and qualifications of a relevant teacher with the qualities of a perceptive item writer. Some such paragons do exist and they produce exceptionally good items, but it has to be said that they are rare.

One of the advantages of employing a professional item writer is that such a person is more likely to be able to reproduce items from one testing occasion to the next: parallel tests are notoriously difficult to write, and the understanding of how test items work that professional writers develop is an important ingredient in the production of consistent tests. However, such professional item writers are likely to be less sensitive to the audience being tested, to changes in the curriculum or its implementation, to varying levels of the school or test population, and to other features of the testing context. Doubtless the best solution is to have item-writing teams consisting of professional item writers and suitably experienced teachers.

3.2 Tests versus exercises

When asking, 'What makes a good test writer?', one might just as well ask, 'What makes a good textbook writer?' The design of a test item is very similar to the design of a learning exercise, where learners are presented with a task or data which they have to cope with in order to develop insights or understanding, and, through feedback (from a teacher, from peers, from introspection and self observation), to develop a capacity to change behaviour and thoughts. Similarly, all test items ask learners to cope with tasks and with data, but in this case it is in order to produce behaviour or language which will give evidence of ability. A test item consists of a method of eliciting behaviour or language, together with a system whereby that behaviour or language can be judged.

We believe, then, that there is no important difference between writing a test item and writing a learning task or exercise. Thus whatever qualities are needed by the designer of an exercise are also

needed by test writers. Perhaps more importantly, the sources of inspiration for exercises can also be used for test writing: test writers, in other words, can and should be as imaginative as possible when thinking about their item types, and one very useful source of ideas is textbooks and other learning materials.

It is interesting that, in our experience, teachers are very reluctant to show outsiders the tests they have written, yet they are usually willing to reveal the sorts of exercises they have developed for classroom use. This may well be because a mystique surrounds test writing in a way that does not apply to exercise writing: tests are thought to be inherently difficult to write. Certainly our experience is that outsiders tend to be much more critical of test items than ever they are of learning exercises, and this can have an inhibiting effect on the test writers.

This reluctance to show one's test items to others may not be due only to the belief that writing tests is difficult. It may also be due to one important difference between tests and exercises that does indeed make test writing more difficult. The fact is that when learners take a test, they do so alone: they receive no assistance from peers or from teachers. Any such assistance would be termed 'cheating'! However, when they do exercises, learners typically expect to receive help from teachers or peers, or at least to feel able to seek help if they need it. Thus, the main difference between a test and an exercise is that with exercises learners get support: with tests, they do not. The effect of this difference is that test items have to be clearer than exercises. The instructions have to be as simple and unambiguous as possible, and the tasks must be familiar to all so that all candidates are measured according to their ability, and not according to their knowledge of what is expected by the test task. Test items, then, have to stand alone in a way that exercises do not. Teachers can compensate for unclear exercises by paraphrasing them, giving examples, or demonstrations, or even simply by skipping those exercises that learners do not understand or are not interested in. The student taking a test has no such possibility, and the test writer therefore has an obligation to ensure that all items are unambiguous. Interestingly, we talk of the validity of a test item: it is very unusual to talk of the validity of an exercise. Yet the concept is applicable to a discussion of learning tasks: tasks that do not enable learners to learn or practise what they are supposed to are invalid. Tests differ from exercises in that they are expected to be valid (and reliable), whereas learning exercises are generally not.

3.3 Where to start?

Item writers have to begin their writing task with the test's specifications (see Chapter 2). This may seem an obvious point, but it is surprising how many writers try to begin item writing by looking at past papers rather than at the specifications. This recourse to past papers is probably due to the fact that many tests lack proper test specifications. There are two problems with trying to replicate or build upon past papers. Firstly, one has to infer test objectives and purposes which are often not readily inferable: objectives and statements of content are implicit in a past paper, and only explicit, usually, in the specifications. Secondly, specifications are much more wide-ranging than a past paper. Any test is of necessity only a sample of what it might have contained. Building a test on previous tests therefore restricts the item writer to what has already been tested. It is normal practice to vary test content, and often test method, for each new test that is written unless there is a requirement to produce a narrowly parallel test, which is certainly not the case for achievement tests and is normally not the case for proficiency tests. Thus it is essential to refer to test specifications in order to ensure as wide a sampling of the potential content and methods as possible.

What to do after consulting the test specifications will depend upon what sort of test one is constructing. If the test is one of language elements – lexis, grammar, and so on in discrete point format – the next step will probably be to consult past papers or some inventory of the content of previous tests in order to avoid the danger of too much duplication of content across tests. Whilst just looking at the content of previous papers can be useful, it is better systematically to classify previous content. Ideally, the test developers will keep a record of the content of all their tests.

Consulting such an inventory will also be a useful second step for item writers who are devising text-based tests, for example of reading and listening, but also possibly of speaking and writing. The record should show what sorts of texts have already been used, and the specifications will indicate the genres, sources, difficulties and so on (see Chapter 2) that are appropriate for the test in question.

For many tests, the item writer's next task is to find appropriate texts. In this case, 'appropriate' means not only texts that match the specifications, but also texts that look as if they will yield suitable items. Not all texts lend themselves to item development, and item writers are well advised to spend some time searching for texts that have promise. Finding suitable texts can be such a problem that item writers often maintain their own 'bank' of texts that can be used in

some future test, and which they are constantly supplementing from their everyday reading. It is sometimes a good idea – and insisted upon by some test developers – to get the approval of the editing or moderating committee for the texts before moving on to develop items or tasks based on them. This is simply expedient: spending time developing items on texts that will ultimately be rejected is both inefficient and depressing.

3.4 Item types

It is important to realise that the method used for testing a language ability may itself affect the student's score. This is called *the method effect*, and its influence should be reduced as much as possible. We are not interested in knowing whether a candidate is good at multiple-choice tests, or can do cloze tests better than other candidates, or finds oral descriptions of a series of pictures particularly difficult. We are interested in finding out about the candidate's grammatical knowledge, or reading ability, or speaking skill. We do not yet know much about test method effect, but as more research looks at how students actually respond to particular test methods, we will begin to understand this effect, or better, these effects, more fully.

Considerable research has, however, been done into some test methods: the cloze technique, and the C-test, for example (see pages 55 and 56). A vast amount of research has been conducted using cloze tests as one of the variables, but rather less has been done to look at what cloze tests measure. What is clear, however, is that different cloze tests measure different things, that is, a test produced by the application of the cloze technique to a text may or may not measure the same thing as a different cloze test on the very same text. This variation appears to be unpredictable and may depend upon which individual words have been deleted. In short, you cannot know in advance what a given cloze test will measure without validating the test in the normal way (see Chapter 8). This means that the method effect of the cloze technique is likely to be quite complex. Nevertheless, there is some evidence that when they take cloze tests, many candidates read in a different way from usual: they read the short amount of context just before the blank, but fail to read the context after the blank. We would argue that this is because of the technique: the existence of blanks at regular intervals tends to induce a form of 'short-text' reading, and many cloze test takers show a lack of attention to the meaning of the wider context that is not shown by their normal reading, when they are indeed context-sensitive.

Similarly, there is evidence that students taking multiple-choice tests can learn strategies for taking such tests that 'artificially' inflate their scores: techniques for guessing the correct answer, for eliminating implausible distractors, for avoiding two options that are similar in meaning, for selecting an option that is notably longer than the other distractors, and so on (see Allan 1992 for an interesting account of a test of test-wiseness, developed in order to identify students who have developed such strategies). There is also evidence from anecdotal accounts of multiple-choice test takers that the test method tends to encourage students to consider alternatives they would not otherwise have considered (see also the discussion of multiple-choice questions in Oller 1979): thus the technique tricks the unwary into making incorrect interpretations they might not otherwise have made.

In addition, it is likely that particular test methods will lend themselves to testing some abilities, and not be so good at testing others. An extreme example is that multiple-choice tests are not suitable for testing a candidate's ability to pronounce a language correctly. Despite suggestions in Lado 1961 and beliefs in Japan to the contrary, Buck 1989 showed clearly that multiple-choice tests of pronunciation do not correlate at all with candidates' abilities to pronounce English phonemes correctly. A less extreme example might be the multiple-choice technique for testing reading ability: it may be easier to control the thought processes of readers with multiple-choice techniques than it is with short-answer-type questions (since the test writer can devise distractors to get candidates to think in certain ways), and this control may be desirable for the testing of inferencing in a foreign language.

Unfortunately, our understanding of the test method effect is still so rudimentary that it is not possible to recommend particular methods for testing particular language abilities. This is perhaps the Holy Grail of language testing.

In the absence of such recommendations, the best advice that can be offered to item writers is: ensure that you use more than one test method for testing any ability. A useful discipline is to devise a test item to cover some desired ability or objective, then to devise another item testing the same ability using a different method or item type. This may lead to increased insight into what certain item types are testing, and ought to lead to a greater understanding of the possibilities of different item types.

In general, the more different methods a test employs, the more confidence we can have that the test is not biased towards one particular method or to one particular sort of learner. In addition, if a series of tests is being developed over a number of years (for example,

end-of-year tests in an institution), we recommend that test developers deliberately vary the methods used so that no one method predominates and becomes predictable (see also Chapter 10). Although we know surprisingly little about how tests affect teaching (see Alderson and Wall 1993 and Wall and Alderson 1993 for a discussion of the issue of washback), it is likely that 'keeping learners guessing' by varying test methods year by year will reduce the predictability of the test's format, and possibly the learning of test-taking strategies for particular test methods.

3.5 Problems with particular item types

In the meantime, even if they do not know the effects of different test methods, item writers need to be aware of the known pitfalls of particular test methods and to learn how to avoid the commonest mistakes in designing certain sorts of test items. Heaton 1988 gives advice on the constructing of different types of test item and how to avoid writing poor items, and there are several publications which give examples of different test types (see, for example, Valette 1977; Hughes 1989 and Weir 1988). We shall not go into test types in any great detail here, therefore, but will just describe some of the most common problems associated with them, starting with objective test types and progressing to more subjective ones.

3.5.1 General problems

There are some problems which apply to all test types, and perhaps the most fundamental is the question of what an item is actually testing. It is very easy with many sorts of test items to test something which is not intended. This item, for example, is supposed to test spelling:

Rearrange the following letters to make English words:

RUFTI	RSOEH	MSAPT
TOLSO	RIEWT	PAHYP

The item *may* be testing spelling, but it is also testing intelligence, ability to do anagrams and, at a pinch, vocabulary. To succeed with this task it may be more important to be able to make the mental leap required, than to be able to spell.

It is very common, unfortunately, especially in high-level proficiency tests, for intelligence to be tested as well as or instead of language. Similarly, background knowledge is frequently tested instead of reading

or listening comprehension. Two examples of such items will be discussed in Section 3.5.2 below.

Another fundamental point is that if one mark is being given for each item, then each item should be independent of the others. Success on one item should not depend on success on another. For example, if it is only possible to answer the second item in a reading comprehension test after correctly answering the first, then a candidate who fails Item 1 will automatically fail Item 2, and will lose two marks rather than one. Some test writers do integrate test items so that success on later items depends on success with earlier ones, but this may lead to problems. This will be discussed in Section 3.5.4 below.

The final point in this general section is that the instructions for all items must be clear. Often students fail a test or an item not because their language is poor, but because they do not understand what they are meant to do. If possible, the language used should be simpler than that in the items themselves, and in some cases the instructions should be written in the candidates' first language. Each new set of items should be preceded by a worked example.

3.5.2 Multiple-choice

The most important requirement of a multiple-choice item is that the 'correct' answer must be genuinely correct. (See Peirce 1992 for interesting comments on this and other problems in the construction of multiple-choice tests of reading.) Although this seems obvious, it is quite possible, especially in reading or listening comprehension tasks, to write an answer that many colleagues would disagree with. Such dubious answers are particularly common in inferencing questions. Every 'correct' answer must therefore be checked with colleagues to avoid problems such as this:

Which is the odd one out?

A	rabbit
B	hare
C	bunny
D	deer

The test writer might have planned that D was the odd one out, but good language learners might have chosen C because 'bunny' is in 'baby-language'.

The other requirement is that item writers must ensure that if the answer key gives just one correct answer, then there is only one correct answer. We are all familiar with items where more than one of the alternatives is correct. Frequently item writers become fixated on a

single answer, and cannot see that one or more of the alternatives are also acceptable. They can only discover this by showing their items to other people.

The following item was written strictly according to the rules given in a beginner's textbook. However, when native speakers were asked which was the correct answer they disagreed with the answer key.

'Why hasn't your mother come?'
'Well, she said she _____ leave the baby.'

A can't
B won't
C couldn't
D mayn't

According to the textbook, C is the correct answer, because of the rules of reported speech. However, many of the native speakers on whom the item was pretested said that A and B were perfectly acceptable, especially in spoken English. In our experience, too rigorous an adherence to what is taught in a textbook may lead to items where there is more than one acceptable answer.

Each wrong alternative should be attractive to at least some of the students. If an alternative is never chosen, then it is wasting everyone's time and might as well not be there. Generally it is a good idea to have at least four alternative answers, so that the chance of a student guessing an answer is only 25%, but if it is impossible to think of a third attractive wrong answer, then it is sensible to have only three alternatives for some items.

Where necessary, multiple-choice items should be presented in context. Often the item writer has a particular context in mind which is not at all obvious to the test takers, and if the context is not given, these students may get the item wrong although they are capable of the language performance required. The presentation of context will often reduce the possibilities of ambiguity, for example:

Select the option closest in meaning to the word underlined:

Come back soon.

A shortly
B later
C today
D tomorrow

The lack of context makes it unclear whether B is really wrong. It would be clearer as follows:

Fill in the blank with the most suitable option:

Visitor: Thank you very much for such a wonderful visit.
Hostess: We were so glad you could come. Come back _____ .

A soon
B later
C today
D tomorrow

This new version also corrects another weakness. In the original version, the correct answer, A, does not fit easily into the stem sentence, as in many contexts it is not common to say 'Come back shortly'. This might worry some of the better students and they might therefore choose a wrong answer. Since there is no direct synonym for 'soon' which would fit into the stem, and since finding synonyms is in any case perhaps unnecessary at this level of English learning, the new version is more appropriate.

The correct alternative should not look so different from the distractors that it stands out from the rest. It should not be noticeably longer or shorter, nor be written in a different style. Heaton 1988: 32 gives the following example when describing poor multiple-choice items:

Select the option closest in meaning to the word underlined:

He began to choke while he was eating the fish.

A die
B cough and vomit
C be unable to breathe because of something in his windpipe
D grow very angry

There are several problems with this item. The most obvious one is that the correct answer, C, is immediately identifiable because it is so much longer than the other alternatives. It looks like a dictionary definition, and any candidate in doubt about the answer would be likely to choose it.

Secondly, distractor B is related to 'choke' semantically, and would therefore be a plausible option for some learners – after all, what does 'closest in meaning' mean? To ensure that B is less 'close in meaning' than the correct answer, the item writer has been forced to provide a 'dictionary definition' so that C is recognisably 'closest in meaning' to the stimulus.

Thirdly, without more context we cannot know for certain whether the man is choking on his food or is very angry. The fact that the sentence is, 'He began to choke while he was eating the fish', rather than the more natural 'He began to choke on a fish bone' or 'He

choked on a fish bone' implies that perhaps he was growing angry instead. Otherwise, why did the sentence say, '... while he was eating'? It is almost as if this is a trick question, and it might confuse the most able students. If the sentence was set in context, the alternatives would be less ambiguous.

Another requirement with multiple-choice questions is that each option should fit equally well into the stem. Heaton 1988: 29 cites the following item where the correct answer, C, does not fit into the stem, because the indefinite article 'a' cannot be used before a word beginning with a vowel:

Someone who designs houses is a _____ .
A designer B builder C architect D plumber

As we mentioned in Section 3.5.1, some items do not test what they are intended to test. This most frequently occurs in comprehension tests, where items may turn out to be testing background knowledge. It is unfortunately easy to write items which can be answered without any reference to the reading or listening passage. For example:

(After a text on memory)
Memorising is easier when the material to be learned is:
A in a foreign language
B already partly known
C unfamiliar but easy
D of no special interest

Even if we do not see the reading passage, it is clear that this is a poor item. Common sense and experience tell us that A is not true, that D is very unlikely, and that B is probably the correct answer. The only alternative which appears to depend on the text for interpretation is C since 'unfamiliar' and 'easy' are both ambiguous.

Such examples are common, even when items have gone through editing procedures. Here is another example from a large national examination, in which five items could be answered without reading the text:

(After a text about trees)
Who gets food from trees?
A Only man
B Only animals
C Man and animals

Whatever the text says, it is surely common knowledge that both humans and animals get food from trees.

This problem of items being independent of the reading or listening passage is not confined to multiple-choice items. It applies to other

objective-type questions, and may also apply to short-answer questions. To make sure that comprehension items are not answerable without reference to the text, item editors should always try answering new comprehension items before they look at or listen to the related text.

A final difficulty that item writers may encounter relates to multiple-choice editing tasks. Students may be given a task in which they are asked to identify the error in a sentence, for example:

A B C
In spite of the rain/the children's teacher/would not allow them/
D E
stay indoors/during playtime.

Here either C or D is the correct answer, depending on whether the students think the error is one of omission or commission. Either of these sentences is correct:

... the children's teacher would not let them stay ...
... the children's teacher would not allow them to stay ...

It is probably sensible to avoid sentences where the error may be one of omission.

3.5.3 Other objective-type items

DICHOTOMOUS ITEMS

True/False or Yes/No items are generally unsatisfactory, as there is a 50% possibility of getting any item right by chance alone. In order to learn anything about a student's ability, it is necessary to have a large number of such items in order to discount the effects of chance. Some item writers reduce the possibility of correct guessing by including a third category such as 'not given' or 'does not say'. This can be useful in a reading comprehension test, but in listening comprehension, especially where the text is only played once, it can be demanding and can lead to student confusion.

MATCHING

By 'matching' we mean items where students are given a list of possible answers which they have to match with some other list of words, phrases, sentences, paragraphs or visual clues. In the following example, the students have to match the four words on the left with

those on the right in order to make other English words. For example, 'car' and 'pet', make 'carpet'.

1	car	A	room
2	cup	B	pet
3	bed	C	dress
4	night	D	board

The disadvantage with this item is that once three of the items have been accurately matched, the fourth pair is correct by default. It is good practice, therefore, to give more alternatives than the matching task requires. The above example would be improved if the students were given a choice of six or seven words in the right hand column.

Note also that it is important in such tasks to make sure that each item in the first column only matches one item in the second.

INFORMATION TRANSFER

Information transfer is used most in reading and listening comprehension tasks. Candidates usually have to transfer material from the text on to a chart, table, form or map. These tasks often resemble real-life activities and are therefore much used in test batteries which try to include authentic tasks. Sometimes the answers consist of just names and numbers, and can be marked objectively. Sometimes they take the form of phrases and short sentences and have to be marked more subjectively. The problems with these latter items are similar to those described in the section below on short-answer questions.

One of the main problems with information transfer questions is that the task can be very complicated. Sometimes the candidates spend so much time working out what should go where in a table that they do not manage to solve what is linguistically an easy problem.

Another problem is that the task may be culturally or cognitively biased. For example, the candidate might be asked to listen to a description of someone's journey through a town and to mark the route on a map. However, students who are unfamiliar with maps or are not good at map-reading are at a disadvantage with tasks of this sort.

ORDERING TASKS

In an ordering task candidates are asked to put a group of words, phrases, sentences or paragraphs in order. Such tasks are typically used to test simple or complex grammar, reference and cohesion, or reading comprehension. Almost all ordering tasks are difficult to construct because it is not easy to provide words or phrases which only make sense in one order. For example, the following question can be

answered in at least two ways:

Put the following words in order to complete the sentence:

She gave _____ .

book her yesterday mother the to

Even more difficult to construct are items where sentences or paragraphs have to be rearranged. For example:

The following sentences and phrases come from a paragraph in an adventure story. Put them in the correct order. Write the letter of each in the space on the right.

Sentence D comes first in the correct order, so D has been written beside the number 1.

A	it was called 'The Last Waltz'	1 D
B	the street was in total darkness	2 ___
C	because it was one he and Richard had learnt at school	3 ___
D	Peter looked outside	4 ___
E	he recognised the tune	5 ___
F	and it seemed deserted	6 ___
G	he thought he heard someone whistling	7 ___

There are at least two ways of ordering this paragraph. The answer key gives 1:D, 2:G, 3:E, 4:C, 5:A, 6:B, 7:F, but 1:D, 2:B, 3:F, 4:G, 5:E, 6:C, 7:A is also acceptable. In this case it is possible to improve the item by adding 'but' to the beginning of phrase G so that the line reads 'but he thought he heard someone whistling'. This makes the second of the two answers the only acceptable one. However, even if it is possible to prepare an item in which the components can only be ordered in one way, it is not always clear what is being tested, and there is always the problem of marking the answers. Say one student makes two mistakes in ordering early in the sequence, but then orders everything else correctly. Should this person get the same mark as someone who has all the ordering wrong? It seems unfair to mark the two the same, but once you start to give different marks for different ordering errors the marking becomes unmanageably complex. Such items are therefore frequently just marked wholly right or wholly wrong, but in that case the amount of effort involved both in constructing and in answering the item may not be considered to be worth it, especially if only one mark is given for the correct version.

EDITING

Editing tests often consist of sentences or passages in which errors have

been introduced which the candidate has to identify. These can take the form of multiple-choice questions, as in Section 3.5.2 above, or can be more open. A common method is to ask students to identify one error in each line of a text, either by marking the text, or by writing a correction beside each appropriate line. The main difficulty with this kind of item is to make sure that there is only one mistake per line.

Some test writers have tried to make the task more realistic by not necessarily having only one error per line, but by asking students to list all the errors while not telling them how many there are. This means that the students may waste much time scouring the text for possible errors, since they can never be satisfied that they have found everything. It also means that the marking is difficult, since if the students miss an early error, or write down a non-existent one, the answers will not line up with the official answer key. At the very least, students should be told how many errors there are. (This applies to most tasks where candidates have to produce a list of some kind.)

GAP-FILLING

'Gap-filling' refers here to tests in which the candidate is given a short passage in which some words or phrases have been deleted. The candidate's task is to restore the missing words. The deletions have been specially selected by the test writer to test chosen aspects of language such as grammar or reading comprehension.

Gap-filling tasks are sometimes based on authentic texts and sometimes on specially written passages. In both cases the major difficulty is to make sure that each gap leads students to write the expected word. Ideally there should only be one correct answer for each gap, but this is generally difficult to achieve. The answer key is therefore likely to have more than one answer for some spaces. For reliability of marking, it is important to reduce the number of alternative answers to the minimum and to ensure that there are no other possible answers which are not listed in the answer key.

Another problem is that candidates may not be able to think of an answer, not because they have poor language but because the word simply does not spring to mind. Here again, this cannot be anticipated by the item writers because they have read the text in its entirety and the missing word therefore seems obvious. Once again it is essential that the test be tried out on colleagues and then pretested.

If many of the gaps are not easily restored, or if marking proves to be a problem, a banked gap-filling task may be the answer. This is a sort of matching task. Each of the missing words or phrases is included in a list which is presented on the same page as the gap-filling text. There are more words in this list than there are gaps in the text, and the

candidate's task is to select the correct word for each gap. There should be only one possible answer for each gap, but candidates should be told that any word in the word list may fit in more than one gap. The words should be listed in alphabetical order.

It is always important to tell students whether each gap is to be filled by one or by more than one word. If more than one word is acceptable, the marking becomes more difficult. If only one word is allowed, then, for clarity's sake, contractions such as 'we'll' should be avoided, and so should hyphenated words.

Sometimes a sentence or phrase is equally good with or without the deleted word. For example:

> It so happened that the man _____ I was following turned out to be extremely fit.

Such items can confuse the students and should be avoided.

CLOZE

Cloze here refers only to tests in which words are deleted mechanically. Each nth word is deleted regardless of what the function of that word is. So, for example, every sixth word might be removed.

As was implied earlier in this chapter, one problem with nth word deletion tasks is that the choice of the first deletion can have an effect on the validity of the test, since once that first word is deleted, all the other deletions automatically follow. Experiments comparing tests based on the same text, but with different initial gaps, and therefore different spaces throughout the passage, have shown that the tests vary in both validity and reliability (Alderson 1978, 1979 and Klein-Braley 1981). Some versions of the test may, for example, have a high proportion of function words deleted, which may be fairly easy for competent language users to restore, and which may distinguish between students of different levels of ability, whereas other versions may have lost a high proportion of content words which may prove to be irretrievable even for native speakers.

Another disadvantage is that an nth word deletion cloze test is not easily amended. If, when it is pretested, some gaps are impossible to complete, how can the test be altered? If the tester decides to reinstate the difficult word and delete another one nearby, then the principle of nth word deletion is being flouted, and if the text is rewritten to make the nth word gap more answerable, the text becomes less authentic.

Marking cloze tests can be difficult since there may be many possible answers for any one gap, and there is often disagreement as to what answers are acceptable. To produce a comprehensive answer key may require wide pretesting of the test and then lengthy discussions on the

appropriacy of different answers. All this will be time-consuming. To avoid this, some testers only accept the exact word that was used in the original text. This naturally leads to lower final scores but does not usually change the students' ranks. However, since it is counter-intuitive to mark someone wrong because they write, say, 'close the door' instead of 'shut the door', it is more common to accept any appropriate answer.

Finally, unless the aim of the cloze test is to test overall language proficiency, as advocated by Oller 1979, such tests may be a wasteful way of testing. Few of the items in any one passage may test the aspects of language with which the tester is concerned. We therefore recommend that, on the whole, test writers construct gap-filling tasks in preference to cloze tests, so that they can delete selected words or phrases in order to test the linguistic features in which they are interested.

C-TEST

C-tests also involve mechanical deletion, but this time it is every second word which is mutilated, and half of each mutilated word remains in the text in order to give the candidate a clue as to what is missing.

C-tests suffer from the same disadvantages as cloze and gap-filling tasks, although the fact that the first few letters of a missing word are given in a C-test text will reduce the number of possible answers for any one gap. Even when the first half of each missing word is provided, though, it is still possible for some answers to be almost unanswerable.

Each blank in the test below must be filled by the second half of a word. If the whole word has an EVEN number of letters, then EXACTLY HALF are missing:

to = t.....; that = th.....; throws = thr.....

If the whole word has an UNEVEN number of letters, ONE MORE THAN HALF are missing:

the = t.....; their = th.....; letters = let.....

Have you heard about a camera that can peer into the ground and 'see' a buried city? Or another th..... can he..... scientists est..... when a vol..... will er.....? Still ano..... that c..... show h..... deeply a bu..... has go..... into fl.....?

The first problem here is that the instructions are too complicated. The task can seem less daunting if the instructions simply tell the candidates that the number of missing letters is shown in each gap. The first gaps in the above example, would then look like this:

Or another th.. can he.. scientists est..... when

The second problem is that the final sentence does not provide enough clues for educated native speakers to be able to complete 'bu.....' and 'fl.....'. This may only be discovered when the test is pretested.

DICTATION

Dictation can only be fair to students if it is presented in the same way to them all, and this generally means having the material on tape, so that not only is it presented in an identical way to all candidates, but the speed of delivery and positioning of pauses can be tested in advance. If the use of a tape recording is impossible, the people who deliver the dictation must be very thoroughly trained.

Dictation can be objectively marked if candidates are asked to write down the original text verbatim, and if the examiner has a system for deciding how marks should be allotted. However, such systems are difficult to devise. For example, if the marking instructions say, 'Deduct one point for each misspelt word and two points for each word that is missing or is not the same as in the original', it is not always clear whether a word is misspelt or just wrong. The same problem occurs even if the marker is told to ignore spelling mistakes.

The other problem with this method of marking dictation is that it is both time-consuming and boring to mark. This means not only that the marking will be expensive but that the markers are likely to make frequent errors. Some test writers avoid this problem by giving a partial dictation in which the candidates are given a copy of the text they are to hear in which words, phrases or sentences have been deleted. The candidates are asked to fill in the gaps as they listen to the text being read.

Some dictation tests do not ask students to write down the words verbatim, but to write down the main points, as a sort of note-taking task. For example, the students might have the programme for a course of study read aloud to them, and might be asked to note down the information they would need if they were following this course. Such a dictation comprises a more authentic listening task than most traditional dictations, but gives rise to problems in marking such as those discussed in the next section.

SHORT-ANSWER QUESTIONS

By 'short-answer questions' we mean items that are open-ended, where the candidates have to think up the answer for themselves. The answers may range from a word or phrase to one or two sentences.

The most important point perhaps to remember when designing

short-answer questions is that the candidates must know what is expected of them. For example, in the following example, it is not at all clear what is wanted:

Rewrite the following sentence, starting with the words provided. The new sentence must be as close as possible to the original.

It was John who saved my life.
If it _____ .

To an item writer who was used to teaching transformations, this would no doubt be a straightforward item, but when it was pretested, most of the students had no idea what they were supposed to write. The task would have been clearer like this:

It was John who saved my life.
If it _____, I would have died.

Sometimes, on the other hand, students think they know what they are supposed to do, when they do not. For example, the following item was supposed to test students' use of the present perfect:

Write two sentences containing 'since'.

Among the students' answers were the following:

Since it is the end of term I am going on holiday.
Since it was raining they stayed at home.

The answers were quite reasonable, but they did not contain the present perfect. If an item writer requires the present perfect, that must be made clear in the instructions. For example:

Complete the following sentence, using the correct form of the verb 'to be':

I _____ here since yesterday.

This could otherwise be tested in a multiple-choice format:

Complete the following sentence.

I _____ here since yesterday.

A am
B was
C will be
D have been

Reading and listening comprehension can be tested using short-answer questions. The answers can be revealing, as they often show

textual misunderstandings which would never have occurred to the test writer. However, the marking of such items is often very difficult since there are frequently many ways of saying the same thing, and many acceptable alternative answers, some of which may not have been anticipated by the item writer. Once again the items must be thoroughly pretested.

3.5.4 Subjectively marked tests

COMPOSITIONS AND ESSAYS

At first sight, writing the prompts for written compositions seems very easy – much easier, for example, than writing multiple-choice questions. All one seems to have to do is write a topic and leave the student to compose an answer. The following kind of prompt is very common:

'Travel broadens the mind.' (J. Smith) Discuss.

There are many disadvantages with this task. One is the problem of terminology. Candidates may not be familiar with the conventions behind the technical use of the word 'discuss', and so will not know what is expected. Test writers must make sure that terms such as 'discuss', and 'illustrate your answer' are understood by all candidates.

The instructions lack information that the candidates need if they are to be able to do justice to themselves.

Candidates need to know how long the essay should be and whether marks will be deducted if it is too short.

They need to know for whom this essay is to be written, so that they can decide whether it should be in a colloquial style as might be used in a letter, or in an academic style similar to that used in a school essay. In the above example, as long as the students are familiar with this technical use of 'discuss', they will know that the essay is to be written in a formal style. Some prompts, however, are less clear.

They need to know how the essay is to be marked. Are the markers looking for fluency or accuracy? Are marks awarded for the structure of the essay, and the ability to present a good argument, or solely for the use of grammar and vocabulary? Candidates need to know all these things in order to decide whether to use easy, well-known structures so as not to be penalised for errors, or whether to take risks because extra marks are awarded for the use of complex and creative language. (The marking of writing tasks of this kind is discussed in Chapter 5.)

Candidates would have a better idea of how to approach this

question if it was presented in the following way:

Write a formal essay for your English teacher saying whether you agree with the saying, 'Travel broadens the mind'.

You should write about 200 to 250 words.
Marks will be given for:
1 structure of the essay, e.g. the use of paragraphs (20%)
2 appropriacy of style (20%)
3 clarity of argument (20%)
4 range of grammar and vocabulary (20%)
5 accuracy of grammar and vocabulary (20%).

A further problem with many writing tasks is that they expect students to have a wide general knowledge. For example:

Describe the legal system in your country.

If the students are not well informed about their legal system, and many will not be, they may not have enough to say to be able to exhibit their level of English proficiency.

Some writing tasks ask students to use creative skills which they may not have. For example:

You are lost in a blizzard. Describe how you try to find your way home.

Other tasks expect students to write interestingly about a subject which might be irrelevant or boring. For example:

Discuss the advantages and disadvantages of living at home during your time at college.

To prevent some of these problems it is better to give students some information before they start writing so that they do not have to be creative. They may be given a short piece of text which sets the scene or provides background information, but it is important that this is short and easy to read so that the candidate does not waste valuable time reading rather than writing, and so that poor readers are not penalised. Some prompts limit the amount of reading required by using a graph or a picture or a series of pictures. In this case it is essential that the graph is easy to understand and that the pictures are clear.

Many tasks, of course, are not as formal as essays and compositions. When a student is asked to write an informal letter or note, it is important that the task should be a natural one. It is not, therefore, advisable to ask candidates to write letters or notes to friends or relations, since they would usually write to such people in their own language. It may be necessary for the item writer to invent a scenario which would require the candidate to write in the foreign language. For

example, the candidate might be asked to write to a friend from abroad, or to leave a note for a landlady.

SUMMARIES

Summaries are used most often to test reading or listening comprehension and writing skills. In some recent tests they have been used for an integrated test of comprehension and writing. Writing summaries may closely replicate many real-life activities, but there are two major problems.

If the candidate writes a poor summary in which some of the main points in the original text are missing, it may be impossible to know whether this is because of poor comprehension or poor writing skills. This does not matter if a single test score is reported for, say, 'summarising a report' and if it is clear that this score is for a combination of reading and writing skills, but it is not reasonable to give the candidate two scores, one for reading and one for writing.

Marking a summary is not easy. Some examiners just give one mark for each main point that the student has written down, regardless of grammar and style. This sounds straightforward but it is not. Identifying the main points in a text is itself so subjective that the examiners may not agree as to what the main points are. The problem is intensified if the marking includes some scheme where, say, main points each get two marks, and subsidiary points get one. If the marking scheme also tries to account for features such as accuracy, fluency and appropriacy, the marking becomes unduly complex.

Some examiners resolve this problem by presenting the original text alongside a summary of it in which key words and expressions are missing. The candidates then have to fill in the missing words in the summary. A well-designed summary task of this kind is a very efficient way of testing reading comprehension, but because there are usually many possible alternative answers for each gap, marking can be difficult, especially if the test is a large-scale one. To avoid this, some examiners ask the candidates only to use the *exact* word from the original text. This should work, but unfortunately there are always some students who do not follow this instruction and enter appropriate, but not exact, answers in the spaces. If these students get low scores although their comprehension of the text is good, then the test cannot be said to be testing reading comprehension.

A good way of avoiding this problem is to provide a bank of possible words and phrases, as in 'banked gap-filling' above. Such tests are difficult to write, and need much pretesting, but can eventually work well and are easier to mark.

ORAL INTERVIEWS

It is sometimes felt that giving someone an oral interview is a quick and painless way of assessing that person's proficiency in the language. Many people think, for example, that if they have a brief chat with a new arrival at a college they will quickly be able to assess the level of that student's language. However, this is not the case. The conversation may turn on superficial topics which may require only a limited vocabulary and which will not stretch the student's ability to use complex structures. This is not the place to discuss oral interviews in detail, but it should be emphasised that the interview needs to be carefully structured so that the aspects of the test which are considered important are covered with each student, and each student is tested in a similar way. It is not fair to the students if some of them are only required to make simple but appropriate comments, while other equally good ones are forced to use complex language which betrays their inadequacies. Interviewers also need to be trained to put candidates at their ease, to get a genuine conversation going without saying much themselves, to manage to appear interested in each interview and to know how to ask questions which will elicit the language required. Chapter 5 briefly discusses the training of oral interviewers.

INFORMATION-GAP ACTIVITIES

Sometimes one, two or more students are given information-gap tasks to complete. For example, two students might be given slightly different pictures and, without seeing each other's, might be asked to work out what the differences are between them. Or a student might be asked to put questions to the interviewer in order to solve some problem. Such tasks can be enjoyable for the candidates, but they are difficult to construct and have a tendency to elicit only a limited range of language. For example, the questioner may be able to get away with using the perfectly acceptable, 'How about ... ?', for many of the questions. In addition this sort of task can be biased. For example, maps are used in many information-gap tasks, but, as we described above, not all candidates may be familiar with maps. All information-gap tasks should be rigorously pretested.

3.6 Test Editing/Moderating Committees

As we have repeatedly emphasised, no one person can possibly produce a good test, or even a good item, without advice. Being close to the

item, as its designer, the item writer 'knows' what the item is intended to test, and will find it difficult to see that it might in fact be testing either something quite different, or something in addition to what is intended. Knowing what the 'correct' answer is means that the item writer has quite a different view of how students will or should process the item from somebody who does not know what the 'correct' answer is.

It is therefore absolutely crucial in all test development, for whatever purpose, at whatever level of learner ability and however trivial the consequences of failure on the test might be, that some person or persons other than the individual item writer look closely at each item, respond to the item as a student would, reflect upon what abilities are required for successful completion of the item/task, and then compare what he or she thinks the item is testing with what the item writer claims it tests. This form of item review should ideally take place at an early stage in the construction process and need not be a formal matter involving a whole committee. The best items are subject to a number of such informal reviews before they reach their final draft stage.

Once items have been edited into this draft stage, they should then be assembled into a draft test paper/subtest, for the consideration of a formal committee. This committee should include experienced item writers (but not normally those who have produced the items being edited), teachers who are experienced in teaching towards the test or in teaching the target group of learners, and possibly other testing experts, or even subject experts if some form of specific purpose test is being prepared.

The task of this committee is to consider each item and the test as a whole for the degree of match with the test specifications, likely level of difficulty, possible unforeseen problems, ambiguities in the wording of items and of instructions, problems of layout, match between texts and questions, and overall balance of the subtest or paper.

It is especially important that members of this editing committee do not simply read the test and its items: they must take each item as if they were students. This means that, for example, for items testing writing skills they must attempt the actual writing task, or for listening items they must hear the tape and try to answer the questions. For listening tests in particular it is important that committee members do not simply read the tapescript and treat it as a reading comprehension exercise; their taking of the test must replicate the experience of the test takers as closely as possible, and must therefore be done with a tape recording if one is required by the test.

This, of course, means that committee members will need to have devoted sufficient time to taking the test in advance of the editing

meeting – a fact often forgotten by institutions who have on their editing committees busy people who may be unable or not inclined to spend the detailed time they should on taking the test.

The conduct of the committee meeting is of considerable importance. Sufficient time must be set aside for an adequate discussion of each item. In our experience, too many editing meetings spend an inordinate amount of time on the first two or three items, run out of time for the remaining items, and rush through the last three quarters of the test at a breakneck speed just to get through the agenda. In addition, in our experience, committees are much more effective before lunch than after it, and all too many committee members seem to need to leave meetings early in order to catch trains home or go on to other meetings.

An effective editing committee will have a firm chairperson who will ensure that ample time is allocated to the meeting, that no more than is necessary is spent on each item, that each member's views are heard and considered (provided that they are reasoned and reasonable), and that clear decisions are taken by the committee and recorded by the secretary or institutional official.

In addition, it is very important that one person is made responsible for ensuring that the recommendations of the committee are not only recorded but also acted upon and implemented in a revised test, which is then subjected to some form of confirmatory vetting before the test is pretested (see Chapter 4).

Although these precautions may seem excessively bureaucratic, it is our experience that when they are not taken, the resulting test is often at least as flawed as it was before the editing committee meeting.

3.7 Survey of EFL Examination Boards: Questionnaire

One board answered 'Not applicable – oral assessment' to all but two of the questions relating to item writing. To avoid repetition, therefore, this board's responses have been omitted from this chapter. It should, of course, be pointed out that oral assessment does require careful consideration (see page 62) since the nature of the task and the criteria for scoring are important components of test design.

QUESTION 9: Are item or test writers given any further information or guidance? ('Further' means 'in addition to the specifications and sample examination papers referred to earlier in the questionnaire'.)

Most boards said they did give item writers further information, but they gave few details. One board said that 'round-table' setting meetings were chaired by the Chief Examiner, and that items were written with guidance and 'vetted' at the meeting. Another said that the Chief Examiners provided detailed setting procedures for question setters, and one said the guidance was 'mostly verbal in committee and accompanying minutes'. Two of the UCLES respondents said that each item writer received 'Notes for guidance', and the subject officer for the Certificate in English for International Business and Trade (CEIBT) said, 'They attend moderating meetings before they become setters. They work in teams of three – one setter for each paper – with the guidance of a more experienced setter. They have two meetings as a team to research the company material and plan tasks.'

Only one board gave extensive information, including a copy of a sample letter sent to item writers (see Chapter 2, page 35 for details).

QUESTION 11: What criteria are used in the appointment of item/test writers?

The boards varied in their requirements. Five said item writers must have appropriate qualifications, of which one specified university and one EFL/ESL qualifications. Six asked for relevant teaching and examining experience or experience in the relevant subject area, while four expected the item writers to be practising teachers accustomed to preparing their students for the relevant exam. One asked for a strong commitment to a communicative approach to teaching and assessment, and one said that a writer's acceptability would depend upon performance at the item-writing meeting.

QUESTION 12: For what period are item/test writers appointed?

There was a variety of responses, ranging from four boards which appointed their item writers annually, to one which did not appoint writers for a specific period and said that the current writers had been 'producing examinations materials for the last fifteen years, experience which ensures continuity and stability'. Two boards did not appoint writers for a given number of years but for a given number of papers.

QUESTION 13: How far in advance of the date of the exam's administration are item writers first asked to produce their items?

Five boards asked their writers to produce their items about two years before the test's administration, and three aimed for about one year. Of the other boards replying, one said item writing was an 'ongoing activity'; one said, 'There is not necessarily a direct link between