Types of Test Items

Types of test item

Whatever purpose a test or exam has, a major factor in its success or failure as a good measuring instrument will be determined by the item types that it contains.

Direct and indirect test items

A test item is direct if it asks candidates to perform the communicative skill which is being tested. Indirect test items, on the other hand, try to measure a student's knowledge and ability by getting at what lies beneath their receptive and productive skills. Whereas direct test items try to be as much like real-life language use as possible, indirect items try to find out about a student's language knowledge through more controlled items, such as multiple choice questions or grammar transformation items. These are often quicker to design and, crucially, easier to mark, and produce greater scorer reliability.

Another distinction needs to be made between discrete-point testing and integrative testing. Whereas discrete-point testing only tests one thing at a time (such as asking students to choose the correct tense of a verb), integrative test items expect students to use a variety of language at anyone given time - as they will have to do when writing a composition or doing a conversational oral test.

In many proficiency tests where students sit a number of different papers, there is a mixture of direct and indirect, discrete-point and integrative testing. Test designers find that this combination gives a good overall picture of student ability. Placement tests often use discrete-point testing to measure students against an existing language syllabus, but may then compare this with more direct and integrative tasks to get a fuller picture.

Indirect test item types

Although there is a wide range of indirect test possibilities, certain types are in common use:

* Multiple choice questions (MCQs): a traditional vocabulary multiple choice question looks like this:

  The journalist was ___ by enemy fire as he tried to send a story by radio.
  A wronged
  B wounded
  C injured
  D damaged

For many years MCQs were considered to be ideal test instruments for measuring students' knowledge of grammar and vocabulary. Above all this was because they were easy to mark, and since the advent of computers the answer books for these tests can be read by machines, not people, thereby cutting out the possibility of scorer error.

However, there are a number of problems with multiple choice questions. In the first place, they are extremely difficult to write well, especially in the design of the incorrect choices. These 'distractors' may actually put ideas into students' heads that they did not

---

1 Taken from Jeremy Harmer’s *The Practice of English Language Teaching (3rd Edition)*
have before they read them. Second, while it is possible to train students so that their MCQ abilities are enhanced, this may not actually improve their English. The difference between two student scores may be between the person who has been trained in the technique and a person who has not, rather than being a difference of language knowledge and ability. MCQs are still widely used, but though they score highly in terms of practicality and scorer reliability, their 'validity' and overall 'reliability' are suspect.

- **Cloze procedures**: cloze procedures seem to offer us the ideal indirect but integrative testing item. They can be prepared quickly and if the claims made for them are true, they are an extremely cost-effective way of finding out about a testee's overall knowledge.

  Cloze, in its purest form, is the deletion of every nth word in a text (somewhere between every fifth or tenth word). Because the procedure is random, it avoids test designer failings. It produces test items like this:

  They sat on a bench attached 1 ____ a picnic table. Below them they 2 ____ see the river gurgling between overgrown 3 _____. The sky was diamond blue, with 4 ____ white clouds dancing in the freshening 5 _____. They could hear the call of 6 ____ and the buzzing of countless insects. 7 ____ were completely alone.

  Cloze testing seems, on the face of it, like a perfect test instrument, since, because of the randomness of the deleted words, anything may be tested (e.g. grammar, collocation, fixed phrases, reading comprehension), and therefore it becomes more integrative in its reach. However, it turns out that the actual score a student gets depends on the particular words that are deleted, rather than on any general English knowledge. Some are more difficult to supply than others, and in some cases there are several possible answers. Even in the short sample text above it is dear that whilst there is no doubt about items such as 1 and 8, for example, item 4 is less predictable. Different passages produce different results.

  Despite such problems of 'reliability', doze is too useful a technique to abandon altogether, however, because it is clear that supplying the correct word for a blank does imply an understanding of context and a knowledge of that word and how it operates. Perhaps it would be better, therefore, to use 'rational' or 'modified' doze procedures (Alderson 1996: 222) where the test designer can be sure that the deleted words are recoverable from the context. This means abandoning the completely random nature of traditional doze procedure. Instead, every eighth or tenth word is deleted, but the teacher has the option to delete a word to the left or right if the context makes this more sensible.

  Modified doze is useful for placement tests since students can be given texts that they would be expected to cope with at certain levels - thus allowing us to judge their suitability for those levels. They are useful, too, as part of a test battery in either achievement or proficiency tests.

- **Transformation and paraphrase**: a common test item asks candidates to rewrite sentences in a slightly different form, retaining the exact meaning of the original. For example, the following item tests the candidates' knowledge of verb and clause patterns that are triggered by the use of *I wish*:

  "I wish"
I'm sorry that I didn't get her an anniversary present
I wish ______________________________

In order to complete the item successfully the student has to understand the first sentence, and then know how to construct an equivalent which is grammatically possible. As such they do tell us something about the candidates' knowledge of the language system.

• Sentence re-ordering: getting students to put words in the right order to make appropriate sentences tells us quite a lot about their underlying knowledge of syntax and lexico-grammatical elements. The following example is typical:

**Put the words in order to make correct sentences.**

called/ I/ I'm/ in/ sorry/ I wasn't / when / you

Re-ordering exercises are fairly easy to write, though it is not always possible to ensure only one correct order.

There are many other indirect techniques too, including sentence fill-ins (Jan __ to the gym every Tuesday morning), choosing the correct tense of verbs in sentences and passages (I have arrived/arrived yesterday), finding errors in sentences (She noticed about his new jacket), and choosing the correct form of a word (He didn’t enjoy being on the lose __ side). All of these offer items which are quick and efficient to score and which aim to tell us something about a student’s underlying knowledge.

**Direct test item types**

For direct test items to achieve 'validity' and to be 'reliable: test designers need to do the following:

• **Create a 'level playing field':** in the case of a written test, teachers and candidates would almost certainly complain about the following essay question:

*Why was the discovery of DNA so important for the science of the twentieth century?*

since it unfairly favors candidates who have sound scientific knowledge and presupposes a knowledge of twentieth-century scientific history.

However, the following topic comes close to ensuring that all candidates have the same chance of success:

---

Some businesses now say that no one can smoke cigarettes in or even near any of their offices. Some governments have banned smoking in all public places - whether outside or inside. This is a good idea but it also takes away some of our freedom. Do you agree or disagree? Give reasons for answer.

General writing question from The IELTS Handbook
Receptive skill testing also needs to avoid making excessive demands on the student’s general or specialist knowledge. Receptive ability testing can also be undermined if the means of testing requires students to perform well in writing or speaking (when it is a test of reading or listening). In such a situation we can no longer be sure that it is the receptive skill we are measuring.

- **Replicate real-life interaction:**
in real life when people speak or write they generally do so with some real purpose. Yet traditional writing tests have often been based exclusively on general essay questions, and speaking tests often included hypothetical questions about what candidates might say if they happened to be in a certain situation. More modern test writers now include tasks which attempt to replicate features of real life (Weir 1993: 167). They will often look similar to the kind of speaking activities described in Chapter 19.

Tests of reading and listening should also, as far as possible, reflect real life. This means that texts should be as realistic as possible, even where they are not authentic (see Chapter 14, HI). Although there are ways of assessing student understanding (using matching tasks or multiple choice questions) which do not necessarily satisfy these criteria, test items should be as much like real reading and listening as possible.

The following direct test item types are a few of the many which attempt to meet the criteria we have mentioned above:

**SPEAKING**
- an interviewer questioning a candidate about themselves
- ‘information gap’ activities where a candidate has to find out information either from an interlocutor or a fellow candidate. (The role-play on page 279, Example 5, would not need much modification to serve as a test item)
- ‘decision-making’ activities, such as showing paired candidates ten photos of people and asking them to put them in order of the best and worst dressed
- using pictures for candidates to compare and contrast, whether they can both see them or whether (as in many communication games) they have found similarities and differences without being able to look at each other’s material
- role-play activities where candidates perform tasks such as introducing themselves, or ringing a theatre to book tickets

**WRITING**
- writing compositions and stories
- ‘transactional letters’ where candidates reply to a job advertisement, or pen a complaint to a hotel based on information given in the exam paper
- information leaflets about their school or a place in their town
- a set of instructions for some common task
- newspaper articles about a recent event
Types of Test Items
=======================================================================================================

READING
- multiple choice questions to test comprehension of a text
- matching written descriptions with pictures of the items, or procedure, they describe
- transferring written information to charts, graphs, maps, etc. (though special care has to be taken not to disadvantage non-mathematically minded candidates)
- choosing the best summary of a paragraph or a whole text
- matching jumbled headings with paragraphs
- inserting sentences provided by the examiner in the correct place in the text

LISTENING
- completing charts with facts and figures from a listening text
- identifying which of a number of objects (pictures on the test paper) is being described
- identifying which (out of two or three speakers) says what
- identifying whether speakers are enthusiastic, encouraging, in disagreement, or amused
- following directions on a map and identifying the correct house or place

In the interests of ‘reliability’, listening tests are most often supplied on tape to ensure that all candidates have the same opportunities, irrespective of the speakers’ voices, speeds, or expressions. Sometimes, as in the computerized TOEFL test (see the notes at the end of this chapter) candidates work with headphones from an individual computer. Where a group of students listen to the same tape or disk, however, we need to be sure that the material is clearly and easily audible.