Common European Framework of Reference for Languages: Learning, Teaching, Assessment

Language examining and test development

Prepared under the direction of
M. Milanovic (A.L.T.E.)

Language Policy Division
Strasbourg, October 2002

TABLE OF CONTENTS

1.0 GENERAL INTRODUCTION

1.1 The purpose of this guide

This guide is designed to help anyone involved with the preparation of language tests, and particularly those wishing to make use of the Council of Europe's "Common European Framework of Reference for languages: Learning, ,teaching, assessment" (put in reference to CUP published version here as well as in reference section??). The aim has been to make the content of the guide relevant not only to test constructors preparing tests in a more formal context such as state examinations, but also to teachers working on school tests. Achieving the correct balance in trying to meet the needs of these two groups has presented something of a challenge; for this reason, readers are encouraged to consider and implement the advice contained in this guide in the light of their purpose in preparing tests and the amount of time and resources they have available. The focus is largely on matters of *process* rather than those of product, in the belief that suitable products emerge from clear principles and well-designed processes rather than the other way round.

1.2 A communicative view of language

The techniques of language testing in use at any time tend to reflect the view of language and the way it is used at that time. What is being tested and the kind of task or item type chosen as a means of testing can be expected to show the influence of current thinking on what language ability is and what exactly we are doing when we use language in everyday life. Communicative language testing evolved out of a shift in language teaching/learning theory and methodology away from a predominantly structural focus towards one that emphasised the importance of language *in use*.

The Council of Europe's Framework is a natural development from earlier work of the Council. It is based on a number of projects which were highly influential world-wide and gained general acceptance in the language professions. These included the Threshold Level (van Ek, 1975; van Ek and Trim, 1990), a manifestation of the communicative approach which has had a widespread and lasting effect on classroom practice and test design. The Preface to the 1980 edition of *Threshold Level English* recommends a functional approach to language teaching in order to 'convert language teaching from structure-dominated scholastic sterility into a vital medium for the freer movement of people and ideas'; the main focus of this approach is on language in practical use, as it serves the daily personal needs of an adult living in a foreign country.

Threshold Level is not in any sense a course, a syllabus or a comprehensive list of the elements of language a learner at a certain level should know; it is a statement of *objectives*, or an attempt 'to specify how a learner should be able to use a language in order to act independently in a country in which that language is the vehicle of communication in everyday life'. This means that learners need to be given the means not only of doing things like buying milk and getting a car repaired, but also exchanging information and opinions with other people, talking about their likes and dislikes and recounting their experiences. The emphasis is firmly on language as a social instrument, or a way of enabling people to interact with one another. The starting point is the range of situations in which language learners commonly find themselves in a foreign country; the goal is to be able to use language to do whatever is necessary in order to act appropriately in those situations.

1.3 A model-based approach to language testing

Since Threshold was first published, a number of models of communicative competence have been put forward. Perhaps the best known model is the one proposed by Canale and Swain (1981) which subdivided communicative competence into four components - grammatical, sociolinguistic, discourse and strategic. In the late 1980s Bachman (1990) presented his first comprehensive view of communicative language ability (CLA), which clearly grew out of the work of Canale and Swain. He suggested that CLA consists of language knowledge or competence combined with the ability to execute that competence in appropriate language use.

A model of language ability is of importance to the language tester because it provides a useful basis for defining the area of competence to be tested. Having a clear idea of what is being tested is a prerequisite for being able to decide whether or not a test is valid (i.e. whether it actually tests what it claims to test); it also makes it possible to develop practical tools for the item writer or test constructor, such as checklists for test content. The overall purpose of any form of language testing is to sample the language abilities of candidates in such a way that a realistic representation of their degree of skill in using language in non-test situations is provided.

The current Framework also contains a model of language ability. Its essence may be presented as a statement about the nature of communicative competence: **communicative competence** (sociolinguistic, linguistic, pragmatic) is a form of **general competence** that leads to **language activity** (interaction, production,

reception, mediation) using **tasks**, **texts** and **strategies** in four principal **domains** (public, occupational, educational, personal) in which arise **situations**, consisting of **locations**, containing **organisations** that structure interaction, **persons** with definite roles, **objects** (animate and inanimate) that constitute an environment, **events** that take place in it, and **operations** that are performed (see Chapter 4 of the Framework document).

The Framework offers language test designers and those involved in producing examinations the possibility of moving collectively towards a shared language testing system that is motivated by the core values of the Council's own notion of European citizenship, while at the same time allowing them to retain their own testing traditions and to enhance in them whatever conforms to accepted professional practice. This guide is directly concerned with the immediate task facing examiners, namely the creation of a broad range of tests that have a definite location and identity within the Framework and that also conform to European and international standards of test production.

1.4 Other factors influencing language test design

It is important to underline that there are not necessarily any right answers in language testing in an absolute sense. No test method need be intrinsically better or worse than any other. The choice of method is made on the basis of a whole range of factors and in the light of responses to a number of questions. For example:

- is the test a test of general proficiency or does it test mainly what has been learnt in a course?
- how much time is available for the test?
- what level of performance is expected?
- is the aim to spread and rank students?
- are the results to be used diagnostically?

Some of these questions are addressed in later sections of this guide.

While the Framework could be said to provide the necessary theoretical approach to language test design and development, this guide is designed to offer a short, integrated account of the practicalities of test construction that have to be recognised by any test designer in order to develop a 'good' test in the most general sense of the term. The focus of the Framework document is firmly on issues of content, whereas this guide concentrates more on the processes involved in test design and development, using the content elements of the Framework as a point of departure. At regular points throughout the guide, the reader is referred directly to specific

sections of the Framework for more detailed guidance. In general, examiners will find Chapter 3 (Common Reference Levels), Chapter 4 (Language use and the language user/learner), Chapter 7 (Tasks and their role in language teaching), and Chapter 9 (Assessment) particularly useful.

Ultimately, the aim of the test constructor is to match the most appropriate method to the stated purpose of a particular test. In this matching exercise it is necessary to try and balance out the important test qualities of reliability, validity, practicality and impact. The mix of these qualities will depend on the reasons for producing a particular test. For a formal selection test, for example, where important decisions are being made about people's lives, reliability and validity will be a priority. For classroom assessment, however, the concern may be more with practicality and impact. The important point is that the test constructor should be fully aware of all the variables which can be manipulated during test development, and that all decisions should be made in a clear and rational manner.

Users of the Guide who are involved in language test design may like to consider and where appropriate state:
- how far the approaches to language testing currently used in their system reflect a particular view of language and the way language is used
- how far these approaches focus upon assessing linguistic knowledge and/or communicative performance
- to what extent these approaches are related to an explicit model of language ability
- how far the Threshold Level specification and the Council of Europe's Framework offer a theoretical approach to language test design and development
- what is the relative importance of language test design factors such as pedagogic culture, social impact, availability of resources, etc.
- what might be an appropriate balance among the important test qualities of reliability, validity, practicality and impact

2.0 THE TEST DEVELOPMENT PROCESS

It is important and useful to think of the process of test development as cyclical and iterative. This involves feeding back the knowledge and experience gained at different stages of the process into a continuous re-assessment of a given test and each administration of it.

Figure 1 shows an attempt to capture this process in diagrammatic form. The diagram offers a comprehensive blueprint for the stages that may be gone through, beginning from the initial perception that a new test is necessary.

- **perceived need for a new test**
- **planning phase**
- **design phase**
- **development phase**
- **operational phase**
- **monitoring phase**

Not all of these stages are always necessary; whether or not they are all included is a rational decision based on the particular requirements of the test development context.

2.1 The cyclical nature of the development process

Figure 1 emphasises the cyclical nature of the test development process.

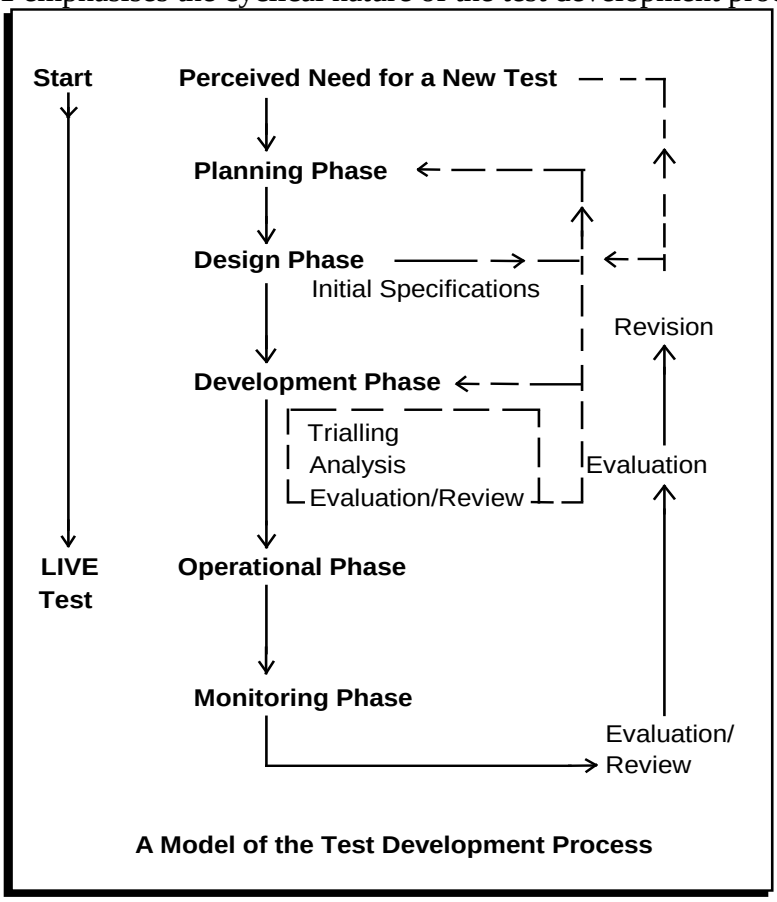

Figure 1 : A test development model

Once the need for a new test has been established, the model involves a **planning phase** during which data on the exact requirements of candidates is collected. In the classroom context, this process may be based on direct personal knowledge of the students and experience of the teaching program. In wider contexts, information may be gathered by means of questionnaires, formal consultation and so on. Whatever the context, the aim will be to establish a clear picture of who the potential candidates are likely to be and who the users of the test results will be.

The planning phase is followed by a **design phase**, during which an attempt is made to produce the initial specifications of a test which will be suitable for the test takers. The specifications describe and discuss the appearance of the test and all aspects of its content, together with the considerations and constraints which affect this. Initial decisions can be made on such matters as the length of each part of the test, which particular item types are chosen, and what range of topics are available for use. At this stage, sample materials should also be written and reactions to these should be sought from interested parties. Even at the level of classroom tests it is always worth showing sample materials to a colleague since another person's reactions can be invaluable in informing the development process.

During the **development phase** the sample materials need to be trialled and/or pretested. This means that students who are at the appropriate level to take the test and who are similar to projected candidates (in terms of age, background, etc.) are given test materials under simulated examination conditions. This phase may involve analysing and interpreting the data provided by candidate scores; useful information can also be gathered by means of questionnaires and feedback reports from candidates and their teachers, as well as video/audio recordings and observations. Decisions can then be made on whether the materials are at the right level of difficulty and whether they are suitable in other ways for use in live tests. Information from trialling also allows fairly comprehensive mark schemes and rating scales to be devised. Even small-scale trialling of classroom or school tests, using just a handful of candidates, can provide valuable information on issues such as the timing allowance needed for individual tasks, the clarity of task instructions, appropriate layout for the response, etc. At this stage it is still possible to make radical changes to the specifications, to the item types used, or to any other aspects of the test which cause concern.

Once the initial phases of **planning**, **design** and **development** have been completed, the test specifications reach their final form, test materials are written, and test papers are constructed. A regular process of administering and marking the test is then set up. This is the **operational phase** (or 'live' phase) during which the test is made available to candidates. (The various stages of this phase are shown in detail in Figure 3 on page 13; the process described here is most applicable to end-of-year school tests, to end-of-course tests in other settings, and to those administered on a wider scale.)

Once a test is fully operational, the test development process enters the **monitoring phase** during which results of live test administrations need to be carefully monitored. This includes obtaining regular feedback from candidates and teachers at schools where the test is used as well as carrying out analyses of candidates' performance on the test; such data is used to evaluate the test's performance and to assess any need for revision. Research may be done into various aspects of candidate and examiner performance in order to see what improvements need to be made to the test or the administrative processes which surround it. Revision of the test is likely to be necessary at some point in the future and any major revision of a test means going back to the planning phase at the beginning of the cycle.

Users of the Guide who are involved in language test development may like to consider and where appropriate state:
- whether a completely new test is required in their situation or whether appropriate revisions can be made to an existing test
- who the potential test candidates are, what their level is, and what their specific requirements are
- how the appearance and content of the test will be affected by local considerations and constraints
- how the administration and marking of the test will be affected by local considerations and constraints
- how adequate trialling/pretesting of the test will be achieved during its development phase
- what methods will be most appropriate for monitoring and evaluating the performance of the live test in the long term
- who the users of the test results will be and how the test results will be interpreted

2.2 Developing test specifications

When the specifications for a new (or revised) test are planned, the underlying aim is always to produce a test which

- is **valid** (i.e. the test should offer an appropriate way of measuring what it claims to measure);
- is **reliable** (i.e. the results produced should be as free as possible from errors of measurement);
- has **impact** (i.e. the effect it has on individuals and on classroom practice should be positive);
- is **practical** (i.e. the demands it makes on the resources of the test developer and the test administrator should be compatible with the resources available).

During planning these factors always need to be kept in mind, and an acceptable balance among them must be achieved.

The first stage of planning should involve a situational analysis. This means looking at the need for a test within the context of the various influences on it which will affect the form it finally takes; the aim of the analysis is to identify the principal **considerations** and **constraints** relevant to the project. These relate to all aspects of what the test must do in order to fulfil its purpose, together with the limitations placed on the test by the circumstances in which it is to be used.

2.2.1 Considerations and constraints

Broadly speaking, the considerations are of two types which can be termed **professional** and **practical**.

Professional considerations concern what exactly it is necessary to test, and include:

- the types of real-life situations in which the candidates will need to use the language;
- the level of performance necessary for those situations;
- the real-life language events which need to be re-created in the testing context;
- the information to be given to users of the test both before and after the test.

Practical considerations are the limitations placed on assessment by factors such as:

- the number of staff and rooms available;
- how many candidates there are;
- how long the test will take;
- the availability of suitably qualified examiners;
- the types of tasks it seems desirable to use;
- the method chosen for reporting scores to candidates;
- the quality control procedures adopted.

Constraints may include:

- the acceptability of the test for all the people involved - candidates, their parents, teachers, owners of schools, etc.;
- the way the test fits into the current system in terms of curriculum objectives and classroom practice;
- the level of difficulty required;
- external expectations of what a test of this kind should be like;
- the availability of resources for test development, test administration and the reporting of results.

This list is not ordered in any specific way nor is it exhaustive. Its purpose here is to emphasise that a good understanding of considerations and constraints is a necessary prerequisite to sensitive and appropriate test design. Chapter 4 of the Framework document provides a useful overview of the many aspects of language use and the language user/learner which need to be considered at this point in test design. They include the context of language use (Section 4.1), the nature of

communicative tasks and purposes (4.3), and the selection of thematic or topic areas (4.2). While the content of Sections 4.1 to 4.4 of the Framework document will be helpful in guiding the situational analysis and in identifying some of the professional considerations which apply, Sections 4.5, 4.6 and Chapter 5 provide a useful basis for determining at a greater level of detail the characteristics of test content for the purpose of developing test specifications.

As Figure 1 shows, once the specifications have been drafted, a first attempt can be made to design the test and to produce sample materials. These can then be trialled and the results analysed. In the light of trialling, some item types or certain types of material may be rejected, and the length of sections of the test or aspects of its administration may be changed. As a result, the specifications may undergo several revisions before they reach the form they are to take for the live test.

It is possible (in the case of a school examination, for example) that the same person will be responsible both for developing the specifications and for writing materials for the live test. However, it must also be possible for people who have not previously been involved in designing or developing the test to get detailed information about it from the specifications. Some people will need this information in order to decide whether to enter students for the test (e.g. if it is a publicly available test). Others may require the information in order to write items for the test; an item writer who has not written for a particular test before and has not been involved in its developmental stages needs a clearly-defined brief to work to; the specifications must go a long way towards providing this.

2.2.2 Content, technical and procedural issues

The specifications which are finally produced should give detailed information on each component or part of the test and will include information about at least three aspects of the test. These are: the **characteristics of test content** (or what is in the test); information on the **technical characteristics** (such as the number of items, sections and so on); and **procedural matters** related to where the test is to be taken and how it is to be graded. Examples of these three aspects are listed below.

Content

- the focus of tasks, e.g. showing detailed comprehension of a text, etc. (see Sections 4.4 and 4.5);
- what is being tested, e.g. use of grammatical rules (see Chapter 5);
- text types used as input (see Section 4.6);
- text sources (see Sections 4.1 and 4.6);
- some indication of topic areas considered suitable for use (see Sections 4.1 and 4.2);
- types of prompts used in tests of oral production (see Sections 4.3 and 4.4);
- types of tasks used in tests of written production (see Sections 4.3 and 4.4)

Technical

- how long the test lasts;
- how many sections it is divided into;
- how many items there are in each section;
- the item types used in each section;
- total and individual length in words of texts used;
- format and length of tasks;
- marks given for each item and total marks available;
- details of weighting;
- where there is a system of examiner marking, details of how the mark scheme is drawn up and teams of examiners co-ordinated;
- details of criteria for assessing free writing tasks and tests of oral production;
- how many examiners or markers are involved, e.g. if double marking is routinely done;
- details of grading procedures and reporting of results.

Procedural

- where and when the test can be taken;
- availability of past papers and/or specimen papers;
- estimated number of hours of study necessary as preparation for test.

All this information helps to give those who need to use the specifications a clear picture of the nature of the materials.

Chapter 4 of the Framework provides a particularly useful reference scheme against which the distinctive features of any test in process of development can be brought into clearer focus. In order to do this, it is necessary first to compile a diagrammatic summary of the test you are working on. Figure 2 offers an example of how the information about an examination made up of five components (or 'papers') can be presented in the form of a grid. Each component of the examination is summarised in terms of what it focuses on, the input provided, and the nature of the expected response.

Paper 1 - Reading	*Test Focus*	*Input*	*Format*
	Understanding structural and lexical appropriacy. Understanding the gist of a written text and its overall function and message. Following the significant points, even though a few words may be unknown. Selecting specific information from a written text. Recognising opinion and attitude when clearly expressed. Showing detailed comprehension of a text.	Section A - discrete sentences. Section B - three or four written texts, covering a range of text types: narrative, descriptive, expository, discursive, informative, etc. Sources include: literary fiction and non-fiction, newspapers, magazines, advertisements, information leaflets, etc.	Section A: twenty-five discrete four-option multiple-choice items. Section B: fifteen four-option multiple-choice items spread across three or four texts.

Paper 2 - Writing	*Test Focus*	*Input*	*Format*
	Using natural and appropriate written language in response to a variety of thematic or situational stimuli.	Four short situational prompts or questions on a range of everyday topics.	Two writing tasks from a choice of five; required length of answer between 120 and 180 words each; the range to include: letters, descriptive/ narrative/ discursive pieces.
Paper 3 - Use of English	*Test Focus*	*Input*	*Format*
	Using English at the word or sentence level, including use of correct structural words and forms; correct and appropriate words and sentences; variety of forms in expressing similar meaning; application of word derivation. Synthesising information in a piece of correct and appropriate extended writing.	Exercises based on short texts and discrete sentences. Some visual input (maps, diagrams, etc.) in directed writing question.	Modified cloze. Transformation exercise. Word formation. Sentence building. Directed writing task.
Paper 4 - Listening	*Test Focus*	*Input*	*Format*
	Understanding the gist of a spoken text and its overall function and message. Following the significant points, even though a few words may be unknown. Selecting specific information from a spoken text. Recognising tone and attitude when clearly expressed. Understanding points of detail in a spoken text.	Three or four authentic or simulated recordings. Sources include: news programmes, news features, conversations, public speeches, announcements, etc.	Three or four tasks, with a total of approximately thirty questions. Task types may include multiple-choice, gap-filling, note-taking, true/false, yes/no, etc.
Paper 5 - Speaking	*Test Focus*	*Input*	*Format*
	Interacting in conversational English in a range of contexts from the everyday to the somewhat more abstract; demonstrating this through appropriate control of fluency, interactive communication, pronunciation at word and sentence level, accuracy and use of vocabulary.	Prompt material including photographs, short texts and visual stimuli. The prompt material may be related to optional background reading texts.	A theme-based conversation between the candidate(s) and an examiner, containing three sections: 1. Talking about a photograph(s) 2. Talking about a short text 3. A communicative activity The interview may be taken singly or in pairs, or in a group of three.

Figure 2 : Information about an examination presented on a grid

Users of the Guide who are involved in drawing up test specifications may like to consider and where appropriate state:
- what type and level of language performance needs to be assessed
- what type of test tasks are necessary to achieve this
- what practical resources are available, e.g. premises, personnel, etc.
- what political, social and/or economic issues are likely to influence test development
- who should be involved in drafting test specifications and developing sample test materials, e.g. in terms of expertise, influence, authority, etc.
- how the content, technical and procedural details of the test will be described in the specifications
- what sort of information about the test needs to be given to users, and how, e.g. a publicly available version of the test specifications

2.3 The production process

The specifications provide a clear definition and detailed description of what must be produced for a test. In this section the focus is on the actual process of production which is likely to consist of five stages:

- **commissioning**
- **vetting and editing**
- **pretesting**
- **analysis and banking of material**
- **question paper construction**

The process of production is illustrated in Figure 3 below but the extent to which these stages are formalised in a given context will depend on who the test is for and how it will be used.

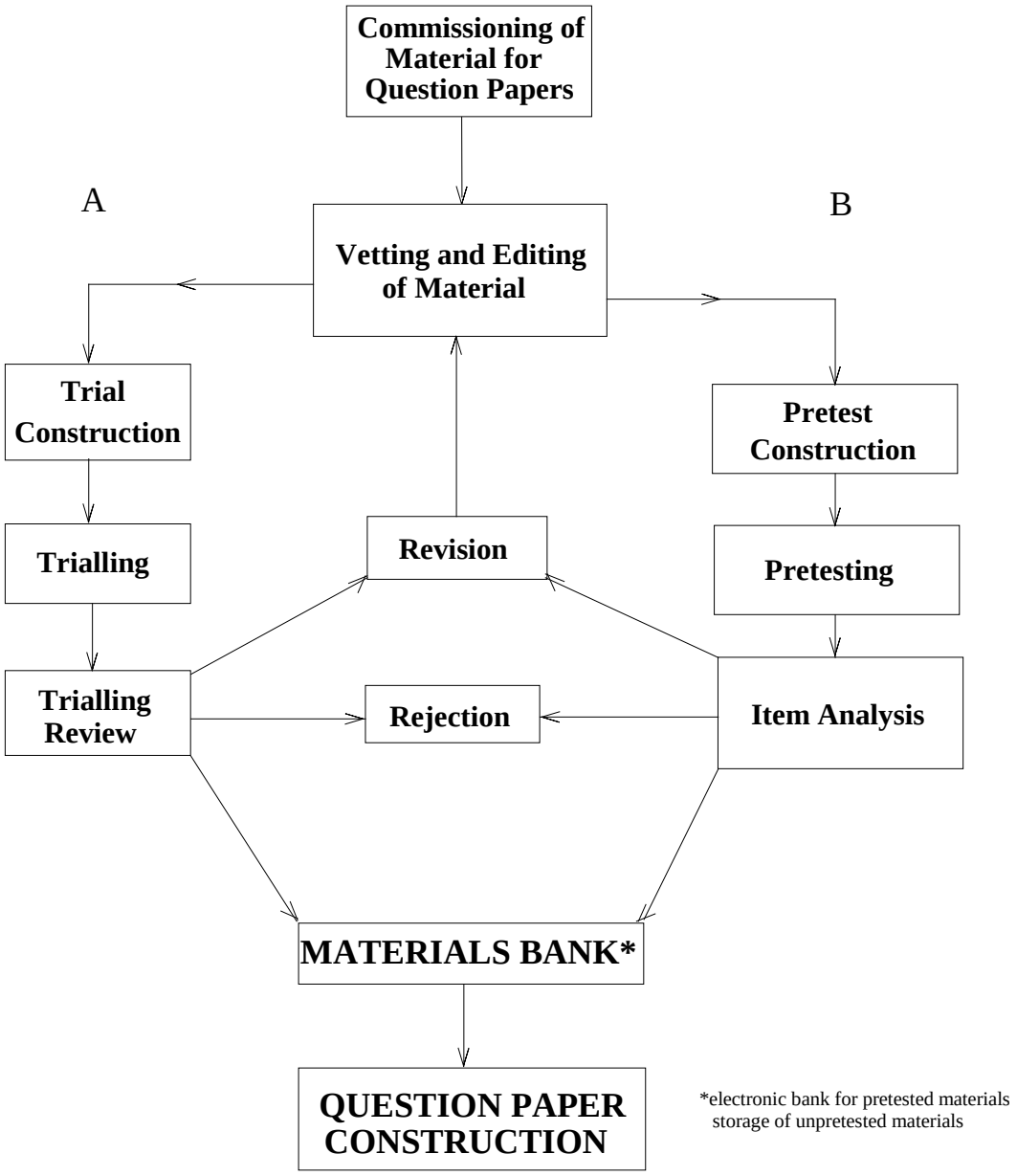

Figure 3 : The operational phase of test production

Figure 3 shows how all material commissioned for question papers passes through an initial vetting and editing stage. At this point the material may follow one of two slightly different routes - A or B - before reaching the point at which it is considered suitable for banking and question paper construction.

In the case of route A, the material is trialled on a fairly small sample population. Although this means that trialling can provide only a very limited statistical analysis, it nevertheless provides valuable information on task effectiveness, level of difficulty, and quality of response; trialling is therefore especially useful in relation to the subjectively marked components of a test, e.g. writing and speaking tasks. In the case of route B, material is pretested on a much larger sample population,

thus offering substantial possibilities for statistical analysis, including item analysis. For this reason, pretesting is particularly appropriate for the objectively marked components of a test.

At all stages of the process, however few or many people are involved, two important principles should be kept in mind:

- **scheduling** (i.e. drawing up realistic project plans, and then meeting deadlines set);
- **record keeping** (i.e. keeping a detailed and accurate account of all decisions and all changes made to materials as they pass through the stages of production).

Appropriate scheduling is essential if materials are to pass efficiently through the different stages of the test production process and ultimately become available for live test use; full and accurate record keeping is vital when any process involving revision and modification is concerned, and when materials may go through several versions.

2.3.1 Commissioning

Commissioning is a term used to describe the process of recruiting people to write test materials. As mentioned previously, it may be that one person (the test developer) is responsible for all the stages of the production process, including the writing of test materials; this is often the case for school-based examinations. However, in other contexts the test developer may commission a number of other people, either employees of the same organisation or outsiders connected with teaching or testing, to take part in selecting or writing texts and in writing items. Sometimes one employee of the organisation which produces a given test is responsible for organising the commissioning and editing stages in this process, and for using the items produced in question paper construction, while other people are involved in the pretesting, analysis and banking stages. A single individual may be responsible for all parts of the test; alternatively, in the case of a test composed of separate parts aimed at reading, writing, listening and speaking for example, different people may be in charge of each component.

Commissioning may follow a regular pattern (e.g. happening twice a year) or it may be done whenever the test developer considers that new materials are needed. Item writers may be asked to submit complete papers, or groups of particular items used in the test.

The aim of the test developer is to receive as high as possible a proportion of material which, after editing, will prove acceptable overall and will find its way into tests and examinations. Part of the test developer's responsibility, therefore, is first to choose suitable people to commission as item writers and then to give them instructions and training which are as clear and helpful as possible. External item writers can often be found among people who have some experience of the test in question; this may be because they prepare students to take the test, or because they are markers or oral examiners. Whether the test developer is working alone or with colleagues or is commissioning external writers, the following points need to be made clear at this stage:

- **Details of the materials required**

  This will include details of the number of texts, tasks and items required.
  In the case of texts, it should be made clear whether items are to be written immediately, or only after prior acceptance of the text. The item writer should provide a key to all items, including correct alternatives.

  A recorded version of listening materials should be requested as well as a written script. This will not be a professionally produced recording; even a simple home-made cassette recording can be very useful for the editing process.

  In the case of a speaking test, it should be made clear whether the item writer is expected to submit visual prompts, or just to indicate what sorts of visual prompts will be needed.

- **Details of the expected presentation of materials**

  Type-written copy is probably the most useful format, and copy may be invited on computer disk as well as on paper. Hand-written copy is less easy to edit and may not be considered acceptable.

  If a whole paper is to be written, the item writer needs to know whether items should be numbered consecutively throughout and the sections run on after each other, or whether each section or exercise should be presented separately, on a new sheet of paper.

  It is helpful if item writers put their name, the date and the name of the test or test paper on each sheet of paper.

  (All the details above can be covered in the guidelines for item writers and are discussed further later in this section.)

- **Details of the deadline by which materials must be submitted**

  It is useful for all item writers to know how their role fits into the overall production schedule; this helps them to understand the importance of keeping to agreed deadlines. At the time of commissioning, it is advisable to give item writers a good idea of when editing of their material will take place; they can then be informed if they are expected to take part in editing, or asked whether they wish to be involved.

- **Details of fees to be paid**

  The terms on which item writers agree to undertake work should be made clear to them at the outset. There may be a fee for accepted materials only, with no payment for any rejected materials; alternatively, a small fee may be paid on initial submission of the materials, to be topped up later for all materials accepted. It may be possible to give a breakdown of rates

payable for various types of item, or simply to give the sum paid for a complete section or test. Teachers who have been asked to write materials for school tests will need to be given enough time within the school timetable to develop the materials.

When writers are commissioned, they will probably be given some of the following documents:

- **specifications**;
- **sample materials** or past papers;
- a set of **instructions or guidelines for item writers** for the specific test or test paper in question.

For large-scale testing it may be necessary to provide writers with additional types of documentation and training, such as:

- a form on which to indicate acceptance of the commission;
- a form on which to indicate that the organisation concerned will own the copyright of the materials to be written;
- a list or lexicon defining the range and level of vocabulary and/or structures to be used;
- a general handbook, giving information about the organisation which produces the tests.

For tests which are commercially produced, a version of the test specifications should be available for public information; while these contain extensive details of the content of an examination, they do not generally include details of test production or of the particular problems which can arise. There may, however, be a fuller version of such a document which is normally confidential, and which includes additional advice and guidance for the item writer. It should contain detailed advice on the **selection and presentation** of materials; this can prevent item writers from wasting time by making their own, possibly mistaken, assumptions about what is acceptable.

Advice on choosing texts

In accordance with the definition given in Section 4.6 of the Framework document, the term "text" is used here to cover any piece of language, whether a spoken utterance or a piece of writing. Advice on choosing texts should refer therefore not only to written texts but also to spoken texts used in listening materials.

It is likely to cover the following points:

- the best sources of texts (e.g. quality newspaper articles, brochures);
- sources less likely to yield acceptable texts (e.g. specialised materials);
- a general warning to avoid bias (e.g. in terms of culture, gender, age, etc.);
- a list of reasons why texts have been rejected in the past.

Reasons for rejecting texts could include:

- too great an assumption of cultural or local knowledge (unless this is being specifically tested);
- unsuitable topics, such as war, death, politics and religious beliefs, which may offend or distress some candidates;
- topics outside the experience of candidates' likely age-group;
- too high a level of difficulty of vocabulary or concept;
- technical or stylistic faults or idiosyncrasies;
- poor editing of the original text.

It may also be possible to give a list of topics which have been covered so well by texts submitted in the past that no more are required.

In the search for suitable texts, Chapters 4 and 7 of the Framework offer considerable help in situating proposed texts within the context of the Council's general notion of language learning. The media listed in Subsection 4.6.2 (voice, telephone, radio, etc.) together with the spoken and written text-types listed in 4.6.3, provide useful checklists and opportunities for diversifying item types.

Advice on presentation

This will probably cover the following points:

- whether typed texts should be double-spaced;
- what information should be given in the heading on each page;
- whether to send in photocopies of original texts;
- which details of text sources to give (e.g. date of publication).

Detailed advice on each question

This is best illustrated with the following example. The task is a modified cloze, designed to focus on words of a structural rather than lexical nature, and the following advice is given to the item writer:

- An authentic text, around 200 words long, is required. It should have a short title. The emphasis is on single structural words. There should not be a heavy load of unfamiliar vocabulary.
- There should be a minimum of sixteen items, more if possible, to allow for selection after pretesting. The first item will be used as an example, and should be numbered (0). Items should test prepositions, pronouns, modifiers, verb auxiliaries, etc. They should be spread evenly through the text, and care should be taken that failing to get one right does not lead automatically to also getting the following one wrong (interdependency of items).

- It is not usually a good idea to gap the first word in a sentence, or to gap a contracted form, as candidates may be confused over whether it counts as one word or two. A gap which leaves a complete grammatical sentence (e.g. gapping the word ‘all’ in the following sentence: *We were informed that all the trains were running late*) should be avoided, as should items which focus on very unusual or idiosyncratic structures.

The standard rubric to be used with this task is also specified for the item writer's benefit.

Having assimilated all the information and advice available, the item writer then has to produce the materials and meet the deadline for submission. Experienced writers of text-based items often gather suitable texts on an ongoing basis from the recommended sources in anticipation of a commission; when the commission arrives, they then select and work on the most promising texts from those already collected. For writing some types of items (e.g. items focusing on grammar or vocabulary), it is useful for the item writer to have a dictionary and thesaurus to hand. When writing listening materials, it is helpful to work with a cassette recorder, so that the test items can be developed directly from the spoken text rather than from the written text on the page.

Many item writers find it useful to try out their materials by asking a colleague or knowledgeable friend not involved in language testing to work through the test task. This helps to identify such faults as typing errors, unclear instructions, incorrect keys and items where the answer is very difficult or where more than one correct answer is possible.

The specifications should also include some form of checklist which the item writer can use to check the text, the items, and the task as a whole, before finally submitting them. The checklist to accompany the modified cloze task described earlier is shown below as an example. If the text, items and task are appropriate, it should be possible to answer each question with "yes".

Text:

Is the text topic accessible/culturally acceptable/etc.?

Is it at the appropriate level of difficulty?

Is the text appropriate for a structurally focused task?

Is it long enough to generate a minimum of sixteen items?

Has a suitable title been included?

Items:

Has the required number of items been generated?

Are the items spread evenly through the text?

Is a good range of language focused on?

Has a check been made that all items are structurally focused?

Is it certain that there are no interdependent items?

Have one or two extra items been included?

Have idiosyncratic items been avoided?

Rubric and key:

Has the rubric been checked?

Has an example (0) been provided?

Has a comprehensive key been provided on a separate sheet?

Before submitting their materials, item writers should check that they have kept a copy of everything; if the originals of texts from newspapers or magazines are being submitted to the test developer, then it is sensible for the item writer to keep photocopies marked up with details of the original source.

2.3.2 Vetting and editing

Once all the item writers who were commissioned have submitted their materials, some preliminary decisions need to be made on which materials should go forward for detailed editing, and which should be rejected immediately or reworked. This stage is sometimes known as **vetting**. It is often undertaken by the test developer, perhaps with the help of another experienced item writer, and is the point at which texts that are clearly unsuitable for any of the reasons given above can be rejected. If texts *without* items have been commissioned, then item writers can be asked at this stage to go ahead and produce items on texts accepted at the vetting stage. Item writers who are asked to submit texts without items should be encouraged to have at least a rough or preliminary outline of the items they intend to write, so that as soon as the text is accepted the items can be supplied as quickly as possible.

Materials that are ready for detailed **editing** can be considered at a meeting attended by a group of item writers and chaired by the test developer or an experienced item writer. The test developer will decide:

- how to group people for the editing sessions;
- which materials each group will consider.

Ideally, materials for editing should be sent out in advance to those who are to attend the editing meeting; this gives everyone an opportunity to work through them beforehand. For text-based items it is worth reading through the items *before* reading the text; this approach helps to highlight any item which can be answered *without* reference to the text (e.g. solely on the basis of common sense or background knowledge). Following this, it is useful to work through the items as if taking the test; this will help to identify, for example, any items in which there is more than one possible correct answer, where the answer is unclear or badly phrased, where there is a distractor so implausible that no candidate who understand it is likely to choose it, or items which are difficult or unclear even to a very proficient user of the language. Reading and listening texts should be checked for their length, suitability of topic, style and level of language. Materials sent out for preparation before the meeting should always be regarded as confidential.

At the editing meeting itself, any problems observed in the materials can be raised and discussed in detail within the group. It is unusual for materials to be accepted exactly as they were submitted and accepted materials are likely to be changed during an editing meeting. Special attention should also be given during the editing meeting to the suitability of rubrics and keys. There is often a lot of discussion about materials and item writers need to be able to accept as well as offer constructive criticism, which can be difficult to do. If an item writer finds it necessary to justify and explain a piece of material to experienced colleagues, then it is likely that the material is flawed in some way. It is useful for the test developer or another person with some degree of authority over the group to be able to make final decisions and decide when there has been enough discussion. In each editing group one person should take responsibility for keeping a detailed and accurate record of all decisions made about materials, showing clearly any changes made at editing. New item writers can be trained in editing by working in a group with more experienced writers. Having more than four or five people in an editing group tends to make the process slow, while fewer than three may not bring in enough variety of points of view.

At the end of the meeting, it is vital that there should be no doubt about what changes were agreed on. For this reason, a clear record of changes made to accepted materials must be kept. Some materials may appear to have potential, but only if they are amended to an extent which could not be done in the course of the meeting. These may be given back to their original writers for further work or may be given to a more experienced writer for revision and further editing. After the meeting, spare and used copies of the edited materials should be destroyed for security reasons. The amended copies of accepted materials are kept by the test developer.

Item writers are entitled to expect some feedback from the test developer on rejected material, especially if they have not been invited to take part in editing, or if they have not been present

during editing of their own materials. This helps item writers to avoid repeating similar mistakes when submitting materials in future.

Users of the Guide who are involved in managing the test production process may like to consider and where appropriate state:
- how the overall test production process will be organised in their situation, e.g. timescales, personnel, procedures, etc.
- who will be commissioned to write test materials
- what level of content knowledge and experience is required
- what training and/or guidance writers will be given
- who will be involved in the process of vetting/editing test materials
- how the vetting/editing process will be managed

2.4 Pretesting and trialling

Pretesting and trialling both involve trying out test materials on a representative sample of the test-taking group to gather various types of information about their performance and measurement characteristics. **Pretesting** is a general term for this sort of activity, but is also used more specifically to refer to occasions when test materials are administered to large groups of test-takers in order to carry out a range of statistical studies on the scores produced. **Trialling** is often used to refer to a form of pretesting involving much smaller groups of test-takers who can provide useful feedback on different performance aspects of the test materials.

The item types which are normally pretested are the more objective item types such as multiple-choice and gap-filling. After the stages of writing and editing, pretesting provides a further, more objective, check on whether a test item works well enough for it to be included in a live test. It is the individual items which are being tested, not the test as a whole, so a pretest paper need not necessarily resemble the actual test for which the material was written, either in length or in composition.

Pretest papers are administered in the form of mock tests under simulated examination conditions to students whose teachers consider them to be at the appropriate language level to take the test. Students benefit from the practice and feedback on their performance which they receive as a result of taking the pretest. In order to carry out the necessary statistical studies and to have confidence in the results, sample sizes of 100-150 or more pretest students are recommended. Trialling is a suitable alternative to pretesting where the latter is not a practical option.

Subjectively marked tests of writing or speaking cannot normally be pretested in the same way as items for which there is a single or limited number of correct answers. In spite of this, some check can be made of how tasks operate before they are used in a live examination. They can be trialled,

again by being administered to students who are at about the correct level for the test, and the answers produced can be marked in line with the normal marking criteria by examiners who are used to marking the live papers. This sort of trialling can show the test developer whether the task was understood by the students, whether it was suitable for their experience and age-group, whether they were provided with enough information to fulfil the task adequately, and whether it gave them the opportunity to show the range of discourse structure, syntactic structure and vocabulary expected of candidates taking an examination at this level.

Both large-scale pretesting and small-scale trialling can be used to gather valuable information on practical aspects of test administration as well as on test-takers' reactions to the test materials.

Statistical analysis of test scores provides the test developer with much useful information about the performance of test items, and can help to prevent the inclusion of poor or faulty items in live tests. It is important to remember, however, that it is always possible for a poor item to produce acceptable statistics; for this reason, the results of this type of analysis should be regarded as only one of the factors determining which materials are used in test papers. An example of an item analysis print-out together with an explanation of what it means is presented in Appendix 1.

Users of the Guide who are involved in test development may like to consider and where appropriate state:
- to what extent it is possible for them to pretest and/or trial test materials in their situation
- what might be the consequences of not pretesting and how these might be addressed
- what type of analysis is to be done on the performance data gathered through trialling/pretesting
- how the results of any analyses will be used, e.g. for test construction purposes, for test writer training, etc.

2.5 Test construction

Test construction is clearly a key activity in the production of question papers to ensure that they meet required standards in terms of difficulty, coverage and content. Approaches to test construction, the nature and level of detail of information collected, and the method of recording this information may differ from test to test. The construction of some tests may be undertaken by a single individual within a specific organisation; other tests may require the involvement of a team of different individuals, some of whom may be internal to an organisation and some of whom may be external consultants.

The test construction stage involves consideration of a number of different variables, all of which must be balanced against one another to produce a test of the required content, coverage and level of difficulty. Certain features of a test may be fixed (e.g. the number of items/tasks to be included), while other features remain flexible (e.g. the topic matter or the variation in accents). If pretest or

trialling data is available, then this information will naturally inform the test construction process. Most tests will normally seek to achieve the correct balance among the following:

- level of difficulty (in terms of the mean difficulty of the test tasks/items and the range of difficulty covered);
- content (in terms of the topics or subject matter);
- coverage (in terms of the representativeness of tasks and testing focus);
- gradedness (in terms of whether the test becomes progressively more difficult);
- item types or test tasks (in terms of the differing cognitive demands they make on test-takers).

Special considerations may apply for certain tests. For example, in a reading test containing several texts and items, a check may need to made to avoid duplication of text-topics or excessive length in terms of total number of words. Similarly, in a listening test, it may be important to check that a balance of male/female voices or of regional accents is maintained.

Once a test has been constructed, it is useful to have it independently vetted. An independent vetter could be someone familiar with the general format of similar tests but who has not been involved with the construction of this particular test; they can be invited to comment on issues relating to suitability of content, continuity of layout/format, etc. If the test is given to someone completely unconnected with the test, then they can provide useful feedback on the clarity of rubrics, layout, etc.

It is important that all decisions made at the test construction stage should be fully and accurately recorded; a template can be used to capture descriptive information, relevant pretest data, the nature of any changes to material, and the rationale for any decisions made. Careful consideration also needs to be given at this stage to checking the rubrics and numbering, and to compiling a comprehensive key or mark scheme.

Where tests form part of a larger examination or series of examinations, the test construction stage should try to take account of the whole examination rather than just individual papers or components. It is important to get an accurate view of the overall balance of a particular examination and to be in a position to compare it with parallel versions at the same level as well as those at different levels and across different administrations. An examination review meeting provides a valuable opportunity for greater communication across papers and across examinations, and enables a coherent overview of examination quality at a point far enough in advance to allow for content or format changes to be agreed if these prove necessary.

Users of the Guide who are involved in constructing tests may like to consider and where appropriate state:
- who will be involved in the activity of test construction in their situation
- which variables need to be considered and balanced against each other, e.g. level of difficulty, topical content, range of item types, etc.
- what will be the role of statistical analyses, e.g. in establishing the mean difficulty and range of the test
- how important statistical analyses will be in relation to other considerations
- whether the constructed test should be independently vetted
- how the constructed test will be matched to parallel forms of the same test or made to fit within a larger series of tests
- how a descriptive profile of the constructed test will be captured, e.g. a record of the topical content, item/task types, measurement characteristics, etc. across the test as a whole

2.6 Issues in Item Writing

In this section we look at some of the issues involved in item writing and offer guidelines that are intended to provide practical help for test writers. The issues addressed in this section concern:

- **task design**;
- **text selection (authenticity, difficulty, etc.)**;
- **choice of item types**;
- **rubrics**;
- **keys, mark schemes and rating scales**.

Once again, Chapters 4 and 7 of the Framework contain a valuable discussion of relevant issues.

2.6.1 Task design

An important general observation is that the type of task designed should be based upon the type of language ability that is being tested and the purpose of testing.

When writing test materials it is vital to achieve an appropriate relationship between the stimulus and the response, and problems are likely to occur if this does not happen. For instance, it is possible to write text-based items which can be answered correctly without the text having been understood. The stimulus may elicit a 'correct' response or answer, but this does not necessarily prove that anything useful has been tested. Similarly, a stimulus may lend itself very easily to a particular item but that particular item may not fit the test purpose.

The difficulty of an item cannot be assumed to be a simple result of the linguistic relationship of the text and the answer. Both stimulus and response have their own linguistic features and the task that bridges them may involve some cognitive complexity in addition to the demands of the language. World knowledge will also play a part as well as other aspects of the Framework model of language

use. When approaching the task of item writing, the writer needs to have a clear idea about the purpose of an item, why that particular item type has been selected and those areas of the test taker's ability that are to be the focus of each item. Sections 7.2 and 7.3 of the Framework discuss in some detail the way in which learner competences, characteristics and strategies interact with task conditions and constraints to affect task performance, in particular task difficulty.

A test may be composed of a number of tasks. The more tightly controlled type of task (such as those used to test reading skills, structural competence, listening and writing at sentence level) is made up of the following components:

- a **rubric** (or instructions for the task);
- some sort of **input to provide a stimulus** (such as a text);
- the **candidate's response** based on items of various types (whether selected or produced);
- a **key or mark scheme**.

A distinction can be drawn between item-based task types and the tasks used in tests of extended writing and speaking, which consist of rubric, input and a response scored against a rating scale or set of criteria as opposed to a key or mark scheme.

2.6.2 Text selection

Item writers are always faced with the task of text selection when preparing materials, particularly for receptive tests of reading or listening and in this section we look at a number of important issues which concern text selection. When selecting texts for a task it is very important to use texts that are suitable for the purpose of testing the particular candidate population concerned. The level of difficulty of the language must be appropriate, and the subject matter must be suitable for the candidates' probable age-group and other aspects of their background. In general, it is better to avoid topics which are beyond candidates' experience, or which might cause distress or offence for some reason. The Framework is an important contribution to any discussion on this topic since it will be increasingly difficult for tests, in spite of their primary concerns with local needs, to avoid the broadening debate on testing in a European context. The most relevant parts of the Framework in connection with text selection are Chapters 4 and 7. Section 4.6 gives a useful list of examples of text-types and the media which carry them; section 4.6.4 considers more closely the nature and function of texts in relation to activities and media.

There are two issues concerning text selection for testing which merit particular attention. One is the issue of authenticity, and the other concerns what makes a text difficult.

Authenticity

Authenticity is an issue affecting choice of texts for teaching as well as for testing and has been a subject for debate since the late 1970s. Is it more appropriate to the candidate's needs for a test to

include a naturally-occurring text (in a test of reading skills, for example) taken from a newspaper or magazine, or a text which has been specially written for the purpose by a test developer or item writer? The newspaper or magazine text may appear 'more authentic', in that it is derived from 'real-life' use of language; it has, after all, been written for native speakers of the language and not just for the purposes of language testing. It could be argued that the goal of language learners is to be able to deal with the texts native speakers have to deal with, and that this is therefore the language they should be exposed to and tested on. It can also be argued that a text written for the sole purpose of testing a certain aspect of language may bear little resemblance to language as it is used by native speakers unconcerned with language testing.

However, it is possible to consider the question of authenticity in broader terms. It has been suggested that authenticity is a consequence of the interaction between a reader and a text, rather than simply a quality of the text alone. With this in mind, a quick look at the range of language use contained in a variety of newspapers and magazines leads to the conclusion that not all written texts are equally relevant to all readers. Who the reader is, the reader's purpose in looking at the text, the writer's purpose and the degree of social and cultural match between reader and text all have a bearing on the nature of the interaction between a reader and a particular text. If there is little match between the factual and cultural knowledge contained in the text and that possessed by the reader, the degree of interaction is likely to be minimal; imagine an elderly opera lover being presented with an article from a teenage rock magazine! As native speakers, we exercise a degree of choice in paying attention to those texts which are relevant to our needs and interests, and avoiding those which are not.

How can the item writer choose texts from naturally-occurring sources such as newspapers and magazines and be sure that they are appropriate for learners of a language who may never have been to any of the countries where that language is spoken, and who cannot be assumed to share any of the social and cultural knowledge of the native speakers for whom the texts were written? It is clearly not enough to cut articles or advertisements out of newspapers and assume that they are useful in language teaching or testing simply because they come from real-life sources rather than out of the heads of test writers. If language learners lack the shared knowledge which the original target readers were assumed to possess, then they are forced back onto a word by word interpretation of the text; this can make the experience of reading artificial and distorted. There has to be a link, however, between test tasks and the non-test language use tasks and situations in which the candidate hopes to be able to use the language and to which the language tester wishes to generalise. There is also the question of face validity, or the degree to which the test materials look convincing to test consumers as a representation of the kind of language use at which they are aiming.

Since the late seventies the notion of authenticity has been extensively explored in order to develop a principled approach to using text in both the teaching and testing of language skills. Widdowson

(1978) and Bachman (1990) conceptualise authenticity as existing at two levels, the **situational** level and the **interactional** level.

i. **situational authenticity**

Situational authenticity refers to the degree to which the test method characteristics of a language task reflect the characteristics of a real life situation where the language will be used; in other words, the extent to which the task is an accurate representation of some language activity which occurs naturally in everyday life.

In designing a situationally authentic task, it is necessary first to identify the critical features that define the task in the target language use domain. It is then possible to design test tasks which have these critical features.

ii. **interactional authenticity**

Interactional authenticity refers to the interaction between test task and test taker; it implies that test writers and developers should:

- make use of texts, situational contexts, and tasks which simulate 'real life' without trying to replicate it exactly;
- attempt to use situations and tasks which are likely to be familiar and relevant to the intended test taker at the given level;
- make clear the *purpose* for carrying out a particular task, together with the intended *audience,* by providing appropriate contextualisation;
- make clear the *criterion for success* in completing the task.

When selecting texts and designing items to accompany them, it is therefore important to give some thought to whether the tasks are situationally authentic and whether the operations candidates are being asked to perform on the texts represent the sorts of processes that might naturally occur in dealing with these texts. Not enough is generally known about how people read or listen to be certain that a test that has been designed is authentic in this sense. However, it is often possible to discern when the match between texts and items is inadequate or very misleading to candidates. Test writers need to be sensitive to the issues involved and be aware of them when preparing materials. The discussion in Section 7.3 of the Framework document is particularly relevant in this regard.

Difficulty of texts

A second important issue to be considered is that of text difficulty and the various characteristics of texts which can influence difficulty. For both written and spoken texts, different factors can affect the degree of difficulty readers and listeners experience when processing texts; this is true for all readers/listeners, whether or not they are in the position of examination candidates.

It seems clear that difficulty is partly related to the linguistic structure of the text. For example, a text composed of short, simple sentences, using the active voice, is likely to be perceived as easier than one composed of long, complex sentences which make frequent use of the passive.

In addition to features of the linguistic structure, other factors which can have an effect on a text's level of difficulty relate to the context in which it is placed. In the case of both spoken and written language, a text may be easier to understand if it addresses the reader or listener directly, rather than putting them in the position of a third party, or 'fly on the wall', simply observing interaction between other characters. The visual support provided by pictures or diagrams (or by video in a listening test) can make a text easier to understand, as can the absence of any time pressure for dealing with the text. If the text is placed in a context which creates an 'information gap', giving candidates a compelling reason for extracting information from the text, this too may help to make it easier; in other words, stimulating the reader/listener's interest can increase accessibility.

Certain content features of a text may also have a bearing on difficulty. In a narrative, for example, a small number of clearly differentiated characters are easiest to deal with. A story about two women and two men who are of different ages, have dissimilar names to one another, and are clearly presented as contrasting characters is likely to be perceived as easier than a story which involves a large number of lightly sketched, minor characters. The sequence of events in a narrative is easier to understand when these are described in the chronological order in which they take place (without the use of flashbacks); if there is a clear link between events - such as that of cause and effect - this also makes the text easier to understand than one containing a series of apparently unconnected events. A listener or reader who already possesses relevant knowledge structures into which the new narrative fits will find it less difficult than someone who lacks these.

Lastly, the type of interaction and the relationship that is set up between the text and the reader or listener is another feature which can affect the degree of difficulty of a text. Extremely formal texts expressing a cold relationship or a very informal, intimate style are both likely to cause more difficulty to readers or listeners than a relatively neutral or moderately informal style.

The issues of difficulty discussed above are of particular importance to an item writer involved in devising listening tasks; the reason for this is that the possibility of looking back over the text as a whole and reviewing the relationships between different parts of the text is not present as it is with a written text. In addition to considering the level of linguistic difficulty (i.e. the complexity of structure and vocabulary used), a writer of listening tasks needs to be aware of the following factors when writing or choosing texts; all of these can affect the amount of processing required over and above the level of simple comprehension, and this in turn can impact on the difficulty level of the text.

- A monologue is the easiest type of speech to follow, especially if the speaker seems to be addressing the listener directly. Two contrasting voices (one male, one female, or one adult and one child) are next easiest. A conversation between two people of the same sex and age, or involving more than two speakers, is often more difficult. Where the speakers have clearly differentiated roles, such as parent and child, the text is easier to follow; on the other hand, a conversation between speakers who have similar roles, for example colleagues of the same sex and similar status discussing a situation at work, is generally more difficult.
- A text involving changes of scene, changes of time reference and a large number of events, will be more difficult than one which is limited to a small number of events, all of which share the same time and setting.
- Where a clear context is established from the beginning of the text, it becomes easier to follow.
- A short text packed with information and accompanied by a proportionately large number of items is difficult for candidates to process, even if the level of language used seems appropriate.
- The inclusion of redundant material in a text, in the form of explanation, rephrasing and repetition, can help to lower the difficulty level of a text.
- Informal language, with its high speed, use of contractions and colloquialisms, its apparent lack of coherent organisation and frequent short turns, often presents a more difficult listening task than more formal language, which tends to be slower, to consist of longer turns and to share more of the features of written language.
- A naturally slow speaker with an expressive voice is easier to understand than someone who speaks fast or in a monotone. It also helps if the speed at which the person speaks is consistent or directly related to the information density of the text.
- Section 7.3 of the Framework document discusses in detail some of the characteristics of a text and of its accompanying task which can lead to increased difficulty.

2.6.3 Choice of item types

One of the most important issues concerning item types is determining which type of item is most appropriate for testing a particular skill in a particular test. This question is normally decided at the test design stage.

The large number of different item types used in language testing can be categorised in various ways. Some are described as objective, in that no human judgement is required in marking them; others demand a constructed response and subjective marking methods. Some are based on receptive skills while others test productive skills. Some are text-based while others are free-standing or discrete. Although some item types are more frequently used than others, it would be inappropriate to believe that these are therefore the best ones to use. The most important criterion for measuring the value of an item type is its appropriateness for use in testing language in a particular situation and for a specified purpose. The item type which provides the most direct means of measuring the desired learning outcome tends to be the best item type to choose.

There are a few general rules to follow when constructing any kind of item:

- items should always attempt to test salient information, rather than information which is peripheral or unimportant;
- normal grammatical conventions should be followed;
- when a new item type is used, an example should be provided, unless the procedure is so simple that this is unnecessary;
- wherever items are text-based it must be necessary to read and understand the text in order to arrive at the correct answer; it should not be possible to answer the item correctly by using background or general knowledge only;
- a text-based item should be written in clear and simple language so that those who understand the text do not fail as a result of not understanding the item.
- text-based items may be placed before or after the text; items which aim to test an overview understanding of the text are often placed before the text to encourage a more superficial processing, while items requiring a more detailed reading or asking for conclusions to be drawn tend to be placed after the text.

One way of dividing item types into two broad groupings is to make a distinction between those which expect the candidate to make a choice of response between various options offered, and those which require the candidate to supply the response. These will be referred to respectively as selection items and candidate-supplied response items. The most common form of the selection type of item is the multiple-choice item, although other item types such as true/false and various kinds of matching fall into the same category, since they demand the same kind of response from the candidate. Generally, tests composed of multiple-choice items are regarded as more objective from a marking point of view than those where the candidate has to supply their own response. Ideally, multiple-choice items should not be used unless they have been pretested and analysed. Section 7.3 of the Framework considers many of the issues to consider when selecting from different response formats for test tasks.

It is important to reiterate that one item-type is not in itself more or less useful than another item type. The selection of an appropriate item type depends on the specific aims and priorities of the test provider. It is possible, for example, to test speaking and writing using either item-based tests or whole tasks. Writing and speaking can be subdivided into skills elements labelled ‘grammar’, ‘vocabulary’, ‘spelling’, 'pronunciation', etc.; viewed at the level of these discrete elements, writing or speaking skills could be assessed by means of item-based tests called either ‘writing’ or something like ‘grammar and usage’ or ‘structural competence’. Those writing or speaking skills which involve the organisation of ideas and arguments, interaction, sequencing and the construction of coherent narrative will probably have to be tested by means of tasks which are not generally item-based. The analysis of users’/learners’ general and language competences in Chapter 5,

together with the discussion of communicative language processes in Section 4.5 of the Framework, offer a useful paradigm within which to consider the item type(s) that are most appropriate for use.

2.6.4 Rubrics

The definition of rubric here is 'the instructions given to a candidate on how to respond to a particular input'. These instructions should include how and where the response is to be recorded; for example, whether it is to be by ticking the correct box or by writing a few words, and whether the response is to be recorded on the question paper itself or on a separate answer sheet. Rubrics are important because they tell the candidate what to do and how to do it; for this reason they need to be very carefully written. Section 7.3 of the Framework discusses the importance of the support given within a task, in terms of the conditions and constraints which can be manipulated for both productive and receptive tasks.

The rubric must present as clearly as possible the task which the examiner is setting the candidate. There should be no room for confusion or need for clarification, otherwise this is likely to create anxiety in the candidate; if a candidate is anxious this may impair their performance and affect the reliability and validity of the test. In listening tests, the rubric is very often not only printed on the question paper but also recorded onto the test cassette. In speaking tests consisting of face-to-face interviews the situation is rather different from that in other kinds of testing; instead of a rubric, verbal instructions are given by the interviewer/interlocutor/examiner. There may even be an opportunity for a candidate to ask for clarification of the task in front of him/her and this could justifiably form part of the interaction.

Here is an example of the rubric for an information transfer task:

> For questions 1-8, read the following informal note which you have received from a colleague. Using the information given, complete the formal announcement by writing the missing words in the correct spaces on your answer sheet. The words you need do not occur in the informal note. The exercise begins with an example (0). Use not more than two words in each space.

The key questions to consider when writing a rubric are:

- how clear is it? (i.e. is it possible to misinterpret the nature of the task?);
- how easy is it to understand? (i.e. is the language used at an appropriate level? This is particularly important in language testing at lower levels of proficiency.);
- how adequate is it? (i.e. is *all* the necessary information given?);
- how relevant is it? (i.e. is *only* necessary information given?);
- how consistent are rubrics?;

With regard to the last point, the language of rubrics should be standardised throughout a test, so that the candidate is able, as far as possible, to follow familiar patterns of instruction. Rubrics should also be consistent between versions of the test.

The rubric is an extremely important part of the test task and item writers need to be encouraged to take as much care over writing the rubric for a task as they do over the items. They may find it useful to use a checklist similar to the one below:

- is the rubric consistent with guidelines for the test?
- if the rubric is new to the candidates is at least one clear example included?
- is the language grammatically correct and appropriate to the level of the test? (This means that the level of language used in the rubric must be below the level of language being tested.)
- is the vocabulary within the existing resources of candidates?
- is the language simple and clear?
- is there any superfluous language?
- are there any double negatives?
- is there any room for ambiguity or misunderstanding?
- does the rubric contain all the necessary information and limitations?

Writers are often too 'close' to the material to spot all possible problems; for this reason it is useful to get a colleague to try the item or tasks out.

Important details which may need to be provided in rubrics include:

- exactly where to find the accompanying input (e.g. page numbers);
- how many words to use in the answer;
- whether or not the same answer may be chosen more than once;
- whether or not answers may be written in any order;
- the approximate number of words to produce for writing tasks;
- a clear indication of the extent of any choice among tasks;
- the number of times a listening text will be heard;
- constraints on the degree to which the input material can be used;
- an indication of the criteria for successful completion of the task.

2.6.5 Keys, mark schemes and rating scales

Any test task which makes use of one of the more objective item types should be accompanied not only by an appropriate rubric but also by an accurate key (set of correct answers) or mark scheme; in cases where a task is to be assessed more subjectively, a rating scale, set of task requirements and marking criteria must be provided.

Where there is only one correct answer, as in multiple-choice items and other selection item types, the item writer should always provide a key to the correct answers. For the types of items which

demand production rather than selection, the writer should provide a mark scheme, giving as exhaustive a list as possible of acceptable answers. Writing a mark scheme is a crucial part of the item writing process, because it is often at this point that the strength or weakness of an item becomes obvious, and an item which had appeared to be acceptable has to be re-written or rejected.

A checklist of questions for item writers might include these questions:

- has an appropriate model answer been written?
- if there is more than one possible answer, have all possibilities been included in the key?
- is the key clear and simple to use?
- is there a clear indication of the number of marks to be awarded for each correct answer or part of an answer?
- is there a sufficiently small number of possible answers?
- have any necessary limitations been specified ? (For example: The candidates must choose two from a list of five options - no marks are awarded if more than two are chosen.)

Consideration may also need to be given to whether correct spelling and/or grammar is required for an answer to be marked correct. It should be remembered that a test may well be marked by someone who is not a language specialist; if there is too little restriction on possible answers, then marking may prove problematic because the marker is confused about what constitutes an acceptable or unacceptable answer.

Various methods of arriving at a score for a speaking or extended writing task are possible. Generally, these involve the use of a rating scale, which may break down the skills being assessed in a speaking test into areas such as pronunciation, fluency and accurate use of structure, and mark each on a scale. To help the assessor, a brief description should be provided of a typical performance at each level. The final score for a speaking task or for an essay may be arrived at by giving a mark on each scale, so that a total score for the task is an aggregate of the past scores. Chapter 3 and the appendices of the Framework offer valuable information on the different approaches to developing rating scales and formulating descriptors, and Chapter 9 discusses some of the issues raised by subjective assessment, including the need to accompany descriptors of performance with actual examples of candidates' work corresponding to the scaled marks and descriptors. The Framework's Appendix B - Illustrative Scales of Descriptors - provides an initial example of how it is possible to describe different levels of performance in different communicative tasks and language skills, and Appendices C and D describe projects for the development of descriptors, including the ALTE Can Do project.

It is a good idea for an item writer to write a sample answer to any kind of test item, whether it is demanded or not. Even if the item writer is providing no more than an essay title, it is important to check whether the topic can be dealt with adequately in the number of words specified, and at the

language level expected of candidates. Faults in items of this kind can be picked up through trialling, but every attempt should be made to eliminate them at this earlier stage.

Users of the Guide who are involved in designing test tasks may like to consider and where appropriate state:

- what is the nature of the relationship between test purpose and task design in their situation

- which item types and tasks will be most appropriate for testing the particular language skills of interest

- what sort of advice to give item writers on selecting texts, e.g. likely sources, inappropriate themes, etc.

- what are the linguistic characteristics of texts which might cause difficulty

- what are the cognitive demands of the item types and tasks selected for test purposes

- how far it is desirable to standardise the language of test rubrics

- what types of markscheme and/or rating scale are most appropriate

- how such markschemes will be developed

3.0 EVALUATING TESTS

Test validation is an integral part of the *process model* of test development. The development cycle begins with considerations of function of outcome (i.e. the purpose of the test); this must include considerations of how the test should be used, how relevant and useful the test will be in terms of social consequences and value implications, and the possible effects it might create (including unplanned side-effects).

In order to develop and deliver a high-quality test, appropriate systems and procedures must be established not only for producing the test but also for evaluating it; systems and procedures need to be implemented for both the development and the operational phases of any test, and are designed specifically to:

- validate the test;
- evaluate the impact of the test;
- provide relevant information to test users;
- ensure that a high quality of service is maintained.

It is generally agreed that tests have an impact on educational processes and on society in general. This impact operates on at least two levels, in terms of:

i) education and society in general;
ii) people who are directly affected by tests and their results.

As a point of principle, test developers should operate with the aim that that their tests will not have a negative impact and, as far as possible, should strive to achieve positive impact. In general terms, this can be achieved through the development and presentation of test specifications and detailed syllabus design, together with the provision of professional support programmes for institutions and individual teachers/students.

Positive impact on teaching and learning is an important aspect of impact which operates at both the general and the specific level. Positive educational impact can be achieved through the following practices:

- the identification of suitable experts within any given field to work on all aspects of test development;
- the training and employment of suitable experts to act as question/item writers in test production;
- the training and employment of suitable experts to act as examiners.

It is important to be able to evaluate the educational impact that tests have within the contexts in which they are used and the routine collection of data provides much of the information needed to investigate the impact and usefulness of a given test. It might be desirable to gather data relating to:

- who is taking the test (i.e. profile of the candidates);
- who is using the test results and for what purpose;
- who is teaching towards the test and under what circumstances;
- what kinds of courses and materials are being designed and used to prepare candidates;
- what effect the test has on public perceptions generally (e.g. regarding educational standards generally);
- how the test is viewed by those directly involved in educational processes (e.g. by students, test-takers, teachers, parents, etc.);
- how the test is viewed by members of society outside education (e.g. politicians, businessmen, etc.)

In summary, good practice in testing relies on the adoption of a process model of test development since it is this that provides the necessary conditions for useful tests to be developed and for validation to take place.

Users of the Guide who are involved in test evaluation may like to consider and where appropriate state:
- what systems and procedures will be required in their situation to monitor and evaluate performance of the test once it is operational
- what specific procedures for analysis will be most appropriate
- how the educational and social impact of their test will be assessed
- what systems and procedures will be required to maintain a high quality of service, e.g. establishment of a code of practice
- how relevant information will be provided for their test users, e.g. documentation, provision of professional support programmes, etc.

REFERENCES

ALTE Code of Practice in ALTE Handbook of European Language Examinations and Examination Systems: ALTE 1998

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

Canale, M. and Swain, M. (1981) A theoretical framework for communicative competence. In A. S. Palmer, P. J. Groot and S. A. Trosper (eds) *The Construct Validation of Tests of Communicative Competence*. Washington, DC: TESOL.

Common European Framework of Reference for Languages: Learning, Teaching, Assessment: Cambridge University Press ISBN 0521 80313 6 (Hardback), 0 521 00531 0 (Paperback)

van Ek, J. A. (1975) *The Threshold Level in a European unit/credit system for modern language learning by adults*. Strasbourg: Council of Europe.

van Ek, J. A. (1980) *Threshold Level English*. London: Pergamon Press.

van Ek, J. A. and Trim, J. L. M. (1990) *Threshold Level 1990*. Strasbourg: Council of Europe.

Widdowson, H. G. (1978) *Language Teaching as Communication*. Oxford: Oxford University Press.

FURTHER READING

Alderson, J.C., Clapham, C. & Wall, D. (1995) *Language Test Construction & Evaluation*. Cambridge: Cambridge University Press.

Alderson, J.C. & Hughes, A. (1981) (eds) *Issues in Language Testing*. ELT Documents 111. London: British Council.

Alderson, J. C. and North, B. (1991) (eds) *Language Testing in the 1990s: The Communicative Legacy*. London: Modern English Publications and the British Council.

Alderson J.C., Krahnke, K. & Stansfield, C (1987) (eds) *Reviews of English Language Proficiency Tests*. Washington, DC: TESOL.

Bachman, L.F. & Palmer, A. (1996) *Language Testing in Practice*. Oxford: Oxford University Press.

Barlow, M. (1989) *Formuler et évaluer ses objectifs en formation*. Chronique sociale, 3e tirage, Lyon.

Bolton, S. (1987) *Evaluation de la compétence communicative en langue étrangère*. CREDIF-HATIER, Coll. Lal, Paris.

Carroll, B.J. (1980) *Testing Communicative Performance*. London: Pergamon.

Delorme, C. (1987) *L'évaluation en questions*. ESF éditeur, 2e édition.

Henning, G. (1987) *A Guide to Language Testing*. Cambridge, Mass: Newbury House.

Hill, C. & Parry, K. (1994) *From Testing to Assessment*. Longman.

Lienert, G. A. and Raatz, U. (1994) *Testaufbau und Testanalyse (5. neubearb. und erw. Auflage)*. Weinheim: Beltz, Psychologie Velags Union.

Luissier, D. (1992) *Evaluer les apprentissages dans une approche communicative*. Hachette.

Mager, R. F. (1986) *Comment mesurer les résultats de l'enseignement*. Bordas, Paris.

Underhill, N. (1987) *Testing Spoken Language*. Cambridge: Cambridge University Press.

Weir, C.(1990) *Communicative Language Testing*. Prentice Hall.

Weir, C. (1993) *Understanding & Developing Language Tests*. Prentice Hall.

APPENDIX 1 : ITEM ANALYSIS

Statistical analysis of test scores provides the test developer with much useful information about the performance of individual test items, and can help to prevent the use of poor or faulty items in live administrations. However, it is important to realise that it is always possible for a poor item to produce acceptable statistics; for this reason, the results of this type of analysis are only one of the factors determining which materials are used in examination papers.

Data gathered at pretesting can be analysed using both classical statistics and Rasch analysis. For a classical statistical analysis, software such as MicroCAT is used. This kind of analysis provides information about the performance of individual items, including **item facility**, **item discrimination** and **distractor tallies**.

Item facility

Knowing the facility value of individual items enables the test developer to ensure that test materials are at the right level of difficulty for the test candidates. Facility is expressed as the proportion of correct responses to the item; it can be reported on a scale of 0 to 1, or as a percentage figure.

In the MicroCAT printout shown in Figure 4, a facility value for each item is given in the 'Prop.Correct' column. Item 8, for example, has a facility of 0.38 (i.e. 38% of the pretest candidates gained the mark available for this item). The appropriate level for a test is at the mid-point of the difficulty range, but an acceptable range of item facility might be set from 33 to 67, or from 20 to 80. In fact, the appropriate level may vary from one test to another, depending on the purpose for which the test is being used; a test of proficiency to be given at the end of a course of study may require a different facility level from that needed for an aptitude test.

A test should include some items towards the extreme ends of the range. For some tests a few easy items are located at the beginning of the test in order to allow the candidates to 'warm up'; sometimes these easy items are not counted in the final score.

Items which fall outside the acceptable range for the test are rejected at this stage; they need not be wasted, however. If an item banking system is in place, they may be banked and considered for use in a test at a different level.

Figure 4 Printed output from MicroCat Analysis (Item statistics)

MicroCAT ™ Testing System
Copyright © 1982, 1984, 1986, 1988, 1993 by Assessment Systems Corporation
Item and Test Analysis Program—ITEMAN ™ Version 3.50

Item Statistics					Alternative Statistics					
Seq. No.	Scale -Item	Prop. Correct	Disc. Index	Point Biser.	Alt.	Prop. Total	Endorsing Low	High	Point Biser.	Key
8	2-1	.38	.52	.48	A	.00	.00	.00		
					B	.38	.13	.66	.48	*
					C	.12	.11	.12	-.01	
					D	.49	.74	.23	-.44	
					Other	.01	.00	.00	-.11	
9	2-2	.71	.42	.42	A	.07	.11	.01	-.16	
					B	.11	.18	.04	-.22	
					C	.10	.16	.00	-.22	
					D	.71	.53	.95	.42	*
					Other	.01	.00	.00	-.13	
10	2-3	.68	.56	.56	A	.68	.39	.96	.56	*
					B	.21	.36	.04	-.37	
					C	.03	.08	.00	-.24	
					D	.07	.14	.00	-.22	
					Other	.01	.00	.00	-.13	
11	2-4	.57	.49	.49	A	.18	.28	.08	-.27	
					B	.15	.19	.09	-.12	
					C	.08	.16	.01	-.31	
					D	.57	.33	.83	.49	*
					Other	.01	.00	.00	-.13	
12	2-5	.61	.63	.54	A	.09	.18	.00	-.22	
					B	.20	.28	.03	-.27	
					C	.61	.32	.96	.54	*
					D	.09	.18	.01	-.28	
					Other	.02	.00	.00	-.09	
13	2-6	.81	.35	.48	A	.11	.20	.04	-.29	
					B	.01	.03	.00	-.11	
					C	.81	.61	.96	.48	*
					D	.07	.17	.00	-.34	
					Other	.00	.00	.00		
14	3-1	.93	.19	.39	A	.93	.81	1.00	.39	*
					B	.07	.18	.00	-.39	
					Other	.01	.00	.00	-.03	

Item discrimination

This statistic concerns the item's ability to discriminate between weaker and stronger candidates. More of those whose final score is high should be getting any given item correct than of those whose final score is low. Two main methods of measuring item discrimination are commonly used:

i) **the discrimination index**;

ii) **the point biserial correlation**.

These can be found in the columns headed Disc. Index and Point Biser. on the MicroCAT printout in Figure 4.

i. **Discrimination index**

Once a test has been administered to a number of candidates the candidates can be ranked (or placed in order) by their test scores. Two groups are then extracted from the sample: the top 30% of candidates, known as the high ability group, and the bottom 30% of candidates, known as the low ability group.

The number of candidates in either of these groups is identical and is represented by N. The number of candidates in each group who got the item right is counted to produce:

n_H (the number of candidates in the high ability group who answered the item correctly);

and

n_L (the number of candidates in the low ability group who answered the item correctly).

The discrimination index, d_i, can then be defined as:

$$d_i = \frac{n_H - n_L}{N}.$$

The value d_i can take any value between -1 and +1.

A discrimination index d_i of +1 implies that all the 'good' students are getting this item correct and that all the 'poor' candidates are getting the item wrong.

A discrimination index d_i of -1 implies that all the 'good' students are getting this item incorrect and that all the 'poor' candidates are getting the item correct.

Items with a d_i of 0.30 or greater are normally considered suitable items for that particular group. It should be noted that the discrimination index is linked to abilities of the particular group of candidates concerned. For example, Item 8 has a discrimination index of 0.52 which suggests that it discriminates well between weak and strong candidates in this group; item 14, however, discriminates rather poorly.

ii. **Point biserial correlation**

The point biserial correlation, r_{pb}, is given by the following formula:

$$r_{pb} = \frac{\overline{X_p} - \overline{X_q}}{S_x}\sqrt{pq}$$

where $\overline{X_p}$ is the mean total score for all those candidates who got this item correct

$\overline{X_q}$ is the mean total score for all those candidates who got this item incorrect

p is the proportion of the total number of candidates who got this item correct

q is the proportion of the total number of candidates who got this item incorrect

s_x is the standard deviation of the test scores for all candidates.

In general, items with a value for the point biserial correlation of greater than 0.30 are considered acceptable. When a point biserial correlation appears with a negative value it means that strong candidates failed to choose the correct answer for that item. This may suggest that an option other than the intended one can legitimately be seen as the correct answer; such an option is referred to as a positive distractor. This item cannot then be used in a test, but it may be possible to revise it by removing the positive distractor and then repretest it.

Distractor tallies

Statistical analysis of multiple-choice items will indicate whether or not distractors are functioning adequately, in other words, whether each is plausible enough to attract some candidates, but not so close to the correct answer that more candidates will choose the distractor than choose the key (the correct answer).

A MicroCAT printout sheet will show the proportion of candidates choosing each distractor in the 'Prop.Total' column. Consider, for example, the following analysis of a 4-option multiple choice item for which the key is C:

A	.15
B	.10
C	.63
D	.12

In this case, the statistics show an item where the key and the distractor options are all performing satisfactorily. Ideally, each distractor for an item should be attracting at least 5% of the test-takers, (i.e. each distractor should show a value of 0.05 or above).

For a different item, however, the key is A and the Prop Total column is as follows:

A	.95
B	.04
C	.01
D	.00

It is clear that this item was so easy that almost every candidate answered it correctly, and that one of the distractors (D) was so weak that nobody chose it.

Columns headed ‘Seq.No.’ and ‘Scale-item’ also appear on the printout. 'Seq. No.' refers to the item’s sequence number within the data set; 'Scale-item' refers to the number of the scale that the item was assigned to and the item’s position within that scale. For example, Item 8 in the overall sequence of items in this dataset is the first in a subset of 6 items which have been designated as Scale 2.

It is also possible to get information on the performance of the whole pretest with that particular group of candidates. An example of a printout is attached as Figure 5. The meanings of the terms used under ‘scale statistics’ are as follows:

N of Items	The number of items included in the analysis.
N of Examinees	The number of candidates included in the analysis.
Mean	For dichotomously scored items - the average number of items that were answered correctly; for multi-point items - the average score for examinees included in the sample.
Variance	The spread of scores around the mean score.
Std. Dev.	The square root of the variance.
Skew	The shape of the distribution.
Kurtosis	The peakedness of the distribution.
Minimum	The lowest candidate score.
Maximum	The highest candidate score.
Median	The middle candidate score.
Alpha	The alpha reliability coefficient for each scale ranging from 0.0 to 1.0; this is an index for the homogeneity of a scale and ideally the value should be as close as possible to 1.

SEM	The Standard Error of Measurement which indicates the likely 'error' in a particular score. SEM = SD Ö (1 - r (test)) SEM = standard error of measurement SD = standard deviation r (test) = reliability of test We can be confident that 70% of the scores will lie within one standard deviation of the mean ($\pm$1 SEM), and 95% confident that the scores will lie within 2 standard deviations ($\pm$2 SEM). Example: A student has a score of 67 on a test with a standard deviation of 9 and a reliability coefficient of 0.9. SEM = $9\sqrt{(1-0.9)} = 2.8$ We can be 70% confident that the candidate's score is between 64.2 and 69.8. We can be 90% confident that the candidate's core is between 61.4 and 72.6.
Mean P	The average proportion of correct answers (for dichotomous items only).
Mean Item-Tot.	The average point biserial across all items in the scale (for dichotomous items only).
Mean Biserial	The average biserial correlation across all items on the scale.
Max Score (Low)	The maximum score a candidate could attain and be included in the low ability group (bottom 27%).
N (Low Group)	The number of candidates in the low ability group (bottom 27%).
Min Score (High)	The minimum score a candidate could attain and be included in the high ability group (top 27%).
N (High Group)	The number of candidates in the high ability group (top 27%).

Figure 5 Printed output from Microcat Analysis (Scale statistics)

MicroCAT ™ Testing System
Copyright © 1982, 1984, 1986, 1988, 1993 by Assessment Systems Corporation
Item and Test Analysis Program—ITEMAN ™ Version 3.50
Time: 15.59
Missing-data option: Compute statistics on all available item responses
There were 270 examinees in the data file.

Scale Statistics

Scale:	1	2	3	4
N of Items	5	10	10	10
N of Examinees	270	270	270	270
Mean	3.230	6.633	8.422	8.163
Variance	0.725	3.321	1.755	2.588
Std. Dev.	0.851	1.822	1.325	1.609
Skew	0.047	-0.348	-0.361	-0.709
Kurtosis	-0.491	-0.202	3.043	-0.148
Minimum	1.000	1.000	2.000	3.000
Maximum	5.000	10.000	10.000	10.000
Median	3.000	7.000	9.000	8.000
Alpha	0.091	0.431	0.318	0.499
SEM	0.812	1.375	1.094	1.138
Mean P	0.646	0.663	0.842	0.816
Mean Item-Tot .	0.428	0.406	0.378	0.415
Mean Biserial	0.676	0.547	0.602	0.621
Max Score (Low)	3	6	8	7
N (Low Group)	168	116	115	89
Min Score (High)	4	8	9	9
N (High Group)	102	85	155	132

APPENDIX 2 : GLOSSARY

administration
The date or period during which a test takes place. Many tests have a fixed date of administration several times a year, while others may be administered on demand.

anchor item
An item which is included in two or more tests. Anchor items have known characteristics, and form one section of a new version of a test in order to provide information about that test and the candidates who have taken it, e.g. to calibrate a new test to a measurement scale.

assessor
Someone who assigns a score to a candidate's performance in a test, using subjective judgement to do so. Assessors are normally qualified in the relevant field, and are required to undergo a process of training and standardization. In oral testing the roles of assessor and interlocutor are sometimes distinguished. Also referred to as examiner or rater.

calibrate
In item response theory, to estimate the difficulty of a set of test items.

calibration
The process of determining the scale of a test or tests. Calibration may involve anchoring items from different tests to a common difficulty scale (the theta scale). When a test is constructed from calibrated items then scores on the test indicate the candidates' ability, i.e. their location on the theta scale.

candidate
A test / examination taker.
Also referred to as examinee.

clerical marking
A method of marking in which markers do not need to exercise any special expertise or subjective judgement. They mark by following a mark scheme which specifies all acceptable responses to each test item.

cloze test
A type of gap-filling task in which whole words are deleted from a text. In a traditional cloze, deletion is every nth word. Other gap-filling tasks where short phrases are deleted from a text, or where the item writer chooses the words to be deleted are commonly referred to as cloze tests, for example 'rational cloze'. Candidates may have to supply the missing words (open cloze), or choose from a set of options (multiple choice or banked cloze). Marking of open cloze may be either 'exact word' (only the word deleted from the original text is taken as the correct response) or 'acceptable word' (a list of acceptable responses is given to markers).

common scale
A way of expressing scores of two or more tests on the same scale to allow a direct comparison of results of these tests. The scores of two or more tests can be expressed on a common scale if the raw scores have been transformed through a statistical procedure, e.g. test equating.

component
Part of an examination, often presented as a separate test, with its own instructions booklet and time limit. Components are often skills-based, and have titles such as Listening Comprehension or Composition. Also referred to as subtest.

computerized marking (scoring)
Various ways of using computer systems to minimize error in the marking of objective tests. For example, this can be done by scanning information from the candidate's mark sheet by means of an optical mark reader, and producing data which can be used to provide scores or analyses.

concurrent validity
A test is said to have concurrent validity if the scores it gives correlate highly with a recognized external criterion which measures the same area of knowledge or ability.

construct validity
A test is said to have construct validity if scores can be shown to reflect a theory about the nature of a construct or its relation to other constructs. It could be predicted, for example, that two valid tests of listening comprehension would rank learners in the same way, but each would have a weaker relationship with scores on a test of grammatical competence.

content analysis
A means of describing and analysing the content of test materials. This analysis is necessary in order to ensure that the content of the test meets its specification. It is essential in establishing content and construct validity.

content validity
A test is said to have content validity if the items or tasks of which it is made up constitute a representative sample of items or tasks for the area of knowledge or ability to be tested. These are often related to a syllabus or course.

convergent validity
A test is said to have convergent validity when there is a high correlation between scores achieved in it and those achieved in a different test measuring the same construct (irrespective of method). This can be considered an aspect of construct validity.

criterion-related validity
A test is said to have criterion-related validity if a relationship can be demonstrated between test scores and some external criterion which is believed to be a measure of the same ability. Information on criterion-relatedness is also used in determining how well a test predicts future behaviour.

descriptor
A brief description accompanying a band on a rating scale, which summarizes the degree of proficiency or type of performance expected for a candidate to achieve that particular score.

directed writing task
Refer to definition for guided writing task.

discrete item
A self-contained item. It is not linked to a text, other items or any supplementary material. An example of an item type used in this way is multiple choice.

discrete-point item
A discrete item testing one specific point of e.g. structure or vocabulary, and not linked to any other items. Discrete-point language testing was made popular in the 1960s e.g. by Robert Lado.

discriminant validity
A test is said to have discriminant validity if the correlation it has with tests of a different trait is lower than correlation with tests of the same trait, irrespective of testing method. This can be considered an aspect of construct validity.

discrimination
The power of an item to discriminate between weaker and stronger candidates. Various indices of discrimination are used. Some (e.g. point-biserial, biserial) are based on a correlation between the score on the item and a criterion, such as total score on the test or some external measure of proficiency. Others are based on the difference in the item's difficulty for low and high ability groups. In item response theory the 2 and 3 parameter models estimate item discrimination as the A-parameter.

discursive composition
A writing task in which the candidate has to discuss a topic on which various views can be held, or argue in support of personal opinions.

double marking
A method of assessing performance in which two individuals independently assess candidate performance on a test.

editing
The process by which examination materials submitted by item writers are modified and put into the form in which they will appear on an examination paper.

equivalent forms
Also known as parallel or alternate forms. Different versions of the same test, which are regarded as equivalent to each other in that they are based on the same specifications and measure the same competence. To meet the strict requirements of equivalence under classical test theory, different forms of a test must have the same mean difficulty, variance and covariance, when administered to the same persons. Equivalence is very difficult to achieve in practice.

error of measurement
Refer to definition for standard error of measurement.

examiner
Refer to definition for assessor.

face validity
The extent to which a test appears to candidates, or those choosing it on behalf of candidates, to be an acceptable measure of the ability they wish to measure. This is a subjective judgement rather than one based on any objective analysis of the test, and face validity is often considered not to be a true form of validity. It is sometimes referred to as 'test appeal'.

facility index
The proportion of correct responses to an item, expressed on a scale of 0 to 1. It is also sometimes expressed as a percentage. Also referred to as facility value or p-value.

gap-filling item
Any type of item which requires the candidate to insert some written material - letters, numbers, single words, phrases, sentences or paragraphs - into spaces in a text. The response may be supplied by the candidate or selected from a set of options.

grade
A test score may be reported to the candidate as a grade, for example on a scale of A to E, where A is the highest

grade available, B is a good pass, C a pass and D and E are failing grades.

grading
The process of converting test scores or marks into grades.

guided writing task
A task which involves the candidate in the production of a written text, where graphic or textual information, such as pictures, letters, postcards and instructions, is used to control and standardize the expected response.

impact
The effect created by a test, both in terms of influence on general educational processes, and in terms of the individuals who are affected by test results.

information transfer
A technique of testing which involves taking information given in a certain form and presenting it in a different form. Examples of such tasks are: taking information from a text and using it to label a diagram; re-writing an informal note as a formal announcement.

input
Material provided in a test task for the candidate to use in order to produce an appropriate response. In a test of listening, for example, it may take the form of a recorded text and several accompanying written items.

integrative task
Used to refer to tasks which require more than one skill or subskill for their completion. Examples are the items in a cloze test, an oral interview, reading a letter and writing a response to it.

item
Each testing point in a test which is given a separate mark or marks. Examples are: one gap in a cloze test; one multiple choice question with three or four options; one sentence for grammatical transformation; one question to which a sentence-length response is expected.

item analysis
A description of the performance of individual test items, usually employing classical statistical indices such as facility and discrimination. Software such as MicroCAT Iteman is used for this analysis.

item banking
An approach to the management of test items which entails storing information about items so that tests of known content and difficulty can be constructed. Normally, the approach makes use of a computer database, and is based on latent trait theory, which means that items can be related to each other by means of a common difficulty scale.

item response theory
A group of mathematical models for relating an individual's test performance to that individual's level of ability. These models are based on the fundamental theory that an individual's expected performance on a particular test question, or item, is a function of both the level of difficulty of the item and the individual's level of ability.

key response
a) The correct option in a multiple choice item.
b) More generally, a set of all correct or acceptable responses to test items.

language for specific purposes (LSP)
Language teaching or testing which focuses on the area of language used for a particular activity or profession; for example, English for Air Traffic Control, Spanish for Commerce.

lexis
A term used to refer to vocabulary.

link item
Refer to definition for anchor item.

live test (item)
A test which is currently available for use, and which must for that reason be kept secure.

mark
The outcome of an examination, often expressed as a percentage. Because of adjustments such as heavier weighting for some items, the mark is not always the same as the total score.

marker
Someone who assigns a score to a candidate's responses to a written test. This may involve the use of expert judgement, or, in the case of a clerical marker, the relatively unskilled application of a mark scheme.

marking
Assigning a mark to a candidate's responses to a test. This may involve professional judgement, or the application of a mark scheme which lists all acceptable responses.

mark scheme
A list of all the acceptable responses to the items in a test. A mark scheme makes it possible for a marker to assign a score to a test accurately.

matching task
A test task type which involves bringing together elements from two separate lists. One kind of matching test consists of selecting the correct phrase to complete each of a number of unfinished sentences. A type used in tests of reading comprehension involves choosing from a list something like a holiday or a book to suit a person whose particular requirements are described.

mean
A measure of central tendency often referred to as the average. The mean score in an administration of a test is arrived at by adding together all the scores and dividing by the total number of scores.

measurement
Generally, the process of finding the amount of something by comparison with a fixed unit, e.g. using a ruler to measure length. In the social sciences, measurement often refers to the quantification of characteristics of persons, such as language proficiency.

multiple-choice item
A type of test item which consists of a question or incomplete sentence (stem), with a choice of answers or ways of completing the sentence (options). The candidate's task is to choose the correct option (key) from a set of three, four or five possibilities, and no production of language is involved. For this reason, multiple choice items are normally used in tests of reading and listening. They may be discrete or text-based.

multiple-choice gap-filling
A type of test item in which the candidate's task is to select from a set of options the correct word or phrase to insert into a space in a text.

multiple-matching task
A test task in which a number of questions or sentence completion items, generally based on a reading text, are set. The responses are provided in the form of a bank of words or phrases, each of which can be used an unlimited number of times. The advantage is that options are not removed as the candidate works through the items (as with other forms of matching) so that the task does not become progressively easier.

objective test
A test which can be scored by applying a mark scheme, without the need to bring expert opinion or subjective judgement to the task.

open-ended question
A type of item or task in a written test which requires the candidate to supply, as opposed to select, a response. The purpose of this kind of item is to elicit a relatively unconstrained response, which may vary in length from a few words to an extended essay. The mark scheme therefore allows for a range of acceptable answers.

optical mark reader (OMR)
An electronic device used for scanning information directly from mark sheets or answer sheets. Candidates or examiners can mark item responses or tasks on a mark sheet and this information can be directly read into the computer. Also referred to as scanner.

predictive validity
An indication of how well a test predicts future performance in the relevant skill.

pretesting
A stage in the development of test materials at which items are tried out with representative samples from the target population in order to determine their difficulty. Following statistical analysis, those items that are considered satisfactory can be used in live tests.

prompt
In tests of speaking or writing, graphic materials or texts designed to elicit a response from the candidate.

proof-reading task
A test task which involves checking a text for errors of a specified type, e.g. spelling or structure. Part of the task may also consist of marking errors and supplying correct forms.

question
Sometimes used to refer to a test task or item.

question paper construction
The process of selecting the items which will make up an examination paper, and adding rubrics and an answer key.

Rasch model
A mathematical model, also known as the simple logistic model, which posits a relationship between the probability of a person completing a task and the difference between the ability of the person and the difficulty of the task. Mathematically equivalent to the one-parameter model in item response theory. The Rasch model has been extended in various ways, e.g. to handle scalar responses, or multiple facets accounting for the 'difficulty' of a task.

rating scale
A scale consisting of several ranked categories used for making subjective judgements. In language testing, rating scales for assessing performance are

typically accompanied by band descriptors which make their interpretation clear.

raw score
A test score that has not been statistically manipulated by any transformation, weighting or re-scaling.

real life approach
In language testing, a view that tests should include task types which resemble real life activities as closely as possible. For example, in a real life approach, the content of a test designed to assess whether candidates can cope with an academic course in a foreign language would be based on a needs analysis of the language and language activities typically found on that course.

register
A distinct variety of speech or writing characteristic of a particular activity or a particular degree of formality.

reliability
The consistency or stability of the measures from a test. The more reliable a test is, the less random error it contains. A test which contains systematic error, e.g. bias against a certain group, may be reliable, but not valid.

response
The candidate behaviour elicited by the input of a test. For example, the answer given to a multiple choice item or the work produced in a test of writing.

role play
A task type which is sometimes used in speaking tests in which candidates have to imagine themselves in a specific situation or adopt specific roles.

rubric
The instructions given to candidates to guide their responses to a particular test task.

scale
A set of numbers or categories for measuring something. Four types of measurement scale are distinguished - nominal, ordinal, interval and ratio.

scale descriptor
Refer to definition for descriptor.

script
The paper containing a candidate's responses to a test, used particularly of open-ended task types.

semi-authentic text
A text taken from a real life source that has been edited for use in a test, e.g. to adapt the vocabulary and/or grammar to the level of the candidates.

sentence completion
choosing them from various options given.

sentence transformation
An item type in which a complete sentence is given as a prompt, followed by the first one or two words of a second sentence which expresses the content of the first in a different grammatical form. For example, the first sentence may be active, and the candidate's task is to present the identical content in passive form.

setting
The whole process by which examination materials are produced and papers constructed.

specifications

A description of the characteristics of an examination, including what is tested, how it is tested, details such as number and length of papers, item types used, etc.

structural competence
Structural competence refers to an individual's ability in and knowledge of the grammatical structures of a language.

syllabus
A detailed document which lists all the areas covered in a particular programme of study, and the order in which content is presented.

'table-top' marking
A method of marking examination papers which involves gathering all the markers together to mark for a limited period of time, rather than sending papers out to be marked by people in their own homes.

task
A combination of rubric, input and response. For example, a reading text with several multiple choice items, all of which can be responded to by referring to a single rubric.

test construction
The process of selecting items or tasks and putting them into a test. This process is often preceded by the pretesting or trialling of materials. Items and tasks for test construction may be selected from a bank of materials.

test developer
Someone engaged in the process of developing a new test.

test method characteristics
The defining characteristics of different test methods. These may include environment, rubric, language of instructions, format, etc.

text
A piece of connected discourse, written or spoken, used as the basis for a set of test items.

text-based item
An item based on a piece of connected discourse, e.g. multiple choice items based on a reading comprehension text.

Threshold specification
A detailed description of a particular level of knowledge of English, developed by the Council of Europe. It is estimated that a beginner needs about 375 learning hours to reach this level.

trait
A physical or psychological characteristic of a person, (such as language ability) or the measurement scale constructed to describe this.

trialling
A stage in the development of test tasks aimed at ascertaining whether the test functions as expected. Often used with subjectively marked tasks such as essay questions, which are administered to a limited population.

transformation item
Refer to definition for sentence transformation.

validity
The extent to which scores on a test enable inferences to be made which are appropriate, meaningful and useful, given the purpose of the test. Different aspects of validity are identified, such as content, criterion and construct validity; these provide different kinds

of evidence for judging the overall validity of a test for a given purpose.

vetting
A stage in the cycle of test production at which the test developers assess materials commissioned from item writers and decide which should be rejected as not fulfilling the specifications of the test, and which can go forward to the editing stage.

Waystage level
A specification of an elementary level of foreign language competence first published by the Council of Europe in 1977 for English and revised in 1990. It provides a less demanding objective than Threshold, being estimated to have approximately half the Threshold learning load.

weighting
The assignment of a different number of maximum points to a test item, task or component in order to change its relative contribution in relation to other parts of the same test. For example, if double marks are given to all the items in Task One of a test, Task One will account for a greater proportion of the total score than other tasks.

word formation
An item type where the candidate has to produce a form of a word based on another form of the same word which is given as input.

This glossary has been compiled from the multi-lingual glossary of testing terms produced by the Association of Language Testers in Europe (ALTE) and published by Cambridge University Press in the Cambridge Studies in Language Testing series:
Hardback 0 521 65099 2
Paperback 0 521 65877 2
CD-ROM 0 521 65824 1