ALTE
MATERIALS FOR THE GUIDANCE OF TEST ITEM WRITERS

(1995, UPDATED JULY 2005)

CONTENTS

Module 2
The test production process

Module 3

Item Types
Some issues of item writing
1. Texts
Authenticity
Difficulty of texts
2. Item Types
A few general rules
Multiple choice and other selection types
Discrete point and text based multiple choice
True / false
Gap-filling (cloze) with multiple choice options
Gap-filling with selection from bank
Gap-filling at paragraph level
Matching
Extra word error detection
Candidate supplied response item types
Short answer item
Sentence completion
Open gap-filling (cloze)
Transformation
Word formation
Transformation cloze
Note expansion
Error correction (proof reading)
Information transfer
Non-item-based task types
Writing tasks
Speaking tasks
3. Rubrics
4. Keys, mark scheme and rating scales
5. Exercises

Module 4

Issues in Marking and Scoring

1. Providing a fair result
- How is the item writer involved in the issues related to marking and scoring a test?
- What are the main issues in marking and scoring?

2. Reliability
- In objective tests
- In subjective tests

3. Some issues in marking and scoring objective tests
- Marking the candidate's response
- Computerised marking
- Clerical marking

4. Some issues in marking and scoring subjective tests
- Marking the candidate's response
- How can the assessment of subjective tests be fair?

Tests of Writing
- Methods of marking
- Examiner training
- Multiple marking

Tests of Speaking
- Differences between marking writing and speaking tests
- Examiner training
- Issues related to fairness in speaking test assessment

Appendix A – Formulas for checking test reliability
Appendix B – An example of a Use of English sub-test
Appendix C – An example of a Writing sub-test with sample answers

GENERAL INTRODUCTION TO THE COURSE

Who is this Course for?

This set of materials is a course of study materials to help in training anyone who is involved in any part of the process of developing, writing, administering and reporting the results of tests of a language learned as a foreign language.

In many cases where teachers need to devise progress tests in order to monitor students on the courses they teach, the same person is likely to be involved in every stage of the process, possibly without the involvement of any additional personnel. In other situations, where widely used state-accredited or commercially distributed proficiency or achievement tests are concerned, people may be involved in only a small part of the process, as item writers, perhaps, or as examiners. These materials will be relevant in either case.

How is the Course organised?

To reflect the various stages and issues which make up the testing process and the differing focuses of interest of people involved, the materials are presented in the form of four modules:

Module One deals with the definition of what is to be tested, i.e. models of language ability
Module Two deals with the test production process
Module Three describes item types
Module Four deals with issues in marking and scoring tests.
Each module can be used independently of the others. However, this division into modules does not mean that people, who are only directly involved in, for example, item writing, should limit their knowledge to that area only of the testing process. It is relevant and important for item writers to know about the aspects of language ability and the test production process which underlie the choice of the type of item they are being asked to produce. It is also relevant for writers to know about the particular problems or issues involved in marking that type of item.

The four Modules deal with different stages in or aspects of the process of testing but they are parts of the same whole.

These materials aim to put forward information, ideas and examples of materials related to foreign language testing which are of use to the maximum number of people involved in the process. There is, therefore, an attempt to generalize from the experience gained by working in particular testing organizations to principles which can be extracted from this experience to apply to language testing in the broadest range of different situations. The ideas put forward do not necessarily reflect how things are done in any one organization, but they do represent principles which experience has shown to be sound.

How can the course be used?

The course is suitable for use either for self-access study, or in a group or class situation. Where appropriate, tasks are included and suggested responses are given.

Terminology

There is a great deal of variety in the way in which different testing organizations and individuals refer to parts of tests and aspects of testing. For the purposes of this course, some of the most important and frequently used terms are defined as follows:

COMPONENT	One subtest in an examination. Usually an examination is made up of subtests in the different skills, and components are commonly called by the names of skills. Components are often referred to as 'papers'.
INPUT	The material provided for the candidate to use in order to produce an appropriate response.
ITEM	Each testing point which is given a separate mark or marks. It may consist of, for example, one gap in a written text, or one multiple choice question with three or four options.
RESPONSE	The candidate's written or spoken response to a task. The term 'answer' is often used in this context.
RUBRIC	The instructions given to a candidate on how to respond to a particular input.
SECTION	The parts each component can be subdivided into. Each may be testing a different area of skill or usage.
TASK	A task is defined precisely as a combination of rubric, input and response. For example, a reading text with several multiple choice items, all of which can be answered by referring to a single rubric, can be classified as a task. If another set of items also relates to the same text, but requires a different rubric in order for a response to be produced, then that combination of text, rubric and items constitutes a separate task. The same text can form part of the input to one or more tasks. The term 'question', which is often used in referring to test tasks, whether or not they are presented as grammatical questions, is best avoided because of its ambiguity.
TEXT	Any passage of language, either heard or read, which is used as input in a test, or produced as output by the candidate.

Module 1

MODULE 1

MODELS OF LANGUAGE ABILITY

The techniques of language testing in current use tend to reflect the view of language and language use held at the 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 ideas on what constitutes language ability and what exactly we are doing when we use language in our everyday lives.

The pyschometric-structural approach

The predominant approach to the design of language tests remained until the end of the 70s much as it had been laid down by experts like Robert Lado during the 1960s. Just as the formal structural analysis of language provided the main focus for language teaching materials during this period, the structural syllabus generated by the structural approach in its various forms provided the main source for language test development. The main design principles for language tests of this kind, based on behaviourist psychology and structuralist theories in linguistics, are well known and illustrated in works by Lado (1961), Valette (1967), Harris (1969), and Heaton (1975).

Test items from this time, often referred to as the psychometric-structuralist era (Spolsky, 1975), are characterized, for the most part, by an emphasis on objectivity of marking. This was achieved by using carefully written discrete-point multiple choice items. Even productive language skills such as writing or pronunciation were tested indirectly or receptively using the multiple choice format. The tests that were produced adopted an implicitly hierarchical view of language proficiency in line with the structural linguistic view of the nature of language i.e. that language is built up from phoneme to morpheme to word to sentence. There were few attempts to define language proficiency explicitly, even though Lado went some way towards a definition when he described the process of language acquisition as the internalization of a series of habits of communication. He wrote:

> These habits involve matters of form, meaning and the distribution of layers of structure, namely those of the sentence, clause, phrase, word, morpheme, and phoneme. (1961, p.22)

Critics of the structuralist approach

Although these discrete-point tests dominated in the 60s and 70s, and are still prevalent in some parts of the world today, they had their critics even then. Carroll (1961) noted that a major limitation of discrete-point tests was that they tested only one element of language at a time. He argued that this did not reflect real language use in most cases. He suggested the use of types of tests that focused on the communicative effect of an utterance rather than discrete-point components. Carroll called such tests integrative. He provided a clear statement of what he meant by integrative testing:

> Since the use of language in ordinary situations calls upon all these aspects (of language), we must further recognize that linguistic performance also involves the individual's capability of mobilizing his linguistic competence and performance abilities in an integrated way, i.e. in understanding, speaking, reading or writing in connected discourse (Carroll 1968, p. 58).

Questions such as who is communicating with whom, for what purpose, in what setting, and the extent to which performance is based on underlying skills were not a major concern in language testing, although it would be unfair to say that they played no part at all. The linguistic aspects of language proficiency were easier to isolate and thus easier to test. Moreover, linguistics itself was at this time focusing mostly on the Chomskian approach to the analysis of language that placed its primary focus on the ideal speaker-listener. The role of language as a means of communication was never considered in any depth. Indeed, Lado makes explicit in the following quotation his view that testing language in use as a means of communication would be a process so problematic as to be not worth attempting:

> The situations in which language is the medium of communication are potentially almost infinite. No one, not even the most learned, can speak and understand his native language in all the situations in which it can be used ... even if we could pick only valid situations and even if we could be sure that understanding these situations occurred through the language used, we would still have the problem of the great variety of situations which must be sampled. The elements of the language on the other hand are limited, and it is more profitable to sample these elements than the great variety of situations in which the language is used. (Lado 1961)

Test writers could produce test items which satisfied measurement and linguistic criteria, but did not make any serious attempt to provide a valid external context. The rationale for the construction of test items was that they appeared on a list of structures rather than that they reflected language use in real life situations. However, since much language teaching methodology was equally structural, the tests were considered valid. Their validity might now be questioned on the grounds that the language segments sampled for test items were neither adequate nor authentic and that the relationship between use and usage was left unexplored (Alderson, 1981).

Despite these reservations, the important contributions made to test design during the psychometric-structuralist era must not be forgotten. Contributions such as the emphasis on statistical analysis, reliability and validity, the planning of test content in relation to linguistic structures and the development of the discrete-point multiple-choice item have been of lasting value.

The shift towards 'language in use'

Communicative language testing evolved out of a shift in language teaching/learning theory and methodology away from a predominantly structural focus to one that emphasized the importance of language in use.

This shift of focus began in linguistics and was continued and modified by developments in related fields such as sociolinguistics. Hymes (1967) developed the notion of the speech event, a term used to refer to language activities that are governed by rules of use. He argued that different speech events demand different sets of rules of use and that their structure can be defined by breaking them down into constituent factors such as participant, setting, purpose, topic, channel, etc. Hymes introduced the concept of communicative competence, making the point that competent linguistic performance depends on more than just linguistic knowledge:

> There are rules of use without which the rules of grammar would be useless. Just as rules of syntax can control aspects of phonology, and just as rules of semantics perhaps control aspects of syntax, so rules of speech acts enter as a controlling factor for linguistic form as a whole. (Hymes 1972: 278)

In language teaching such ideas were developed most fully by Munby (1978). His approach is based on the premise that the language to be taught should be related as closely as possible to the learner's immediate and future needs, that the learner should be prepared for authentic communication, and that the language taught should have a high surrender value (Wilkins, 1976). This approach was also developed by van Ek in his Threshold specification, published by the Council of Europe in 1975, and later updated by van Ek and Trim as Threshold Level 1990.

The contribution of Threshold Level

Threshold, as a manifestation of the communicative approach, has had a widespread and lasting effect on classroom practice and on test design. In the Preface to the 1980 edition it is stated that a functional approach to language teaching is recommended in order to 'convert language teaching from structure-dominated scholastic sterility into a vital medium for the freer movement of people and ideas', and the whole focus of this initiative is on language in practical use, as it may serve 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, but it is a statement of objectives, 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. There is an emphasis on language as a social instrument, a way of enabling people to interact with one another. The starting point is the situations in which language learners commonly find themselves in a foreign country, and the goal is to be able to use language to do whatever one needs to in order to act appropriately in those situations.

In the Threshold specification the elements of language are not classified according to structure, but divided into 'functions' and 'notions', which are related to the things people do by means of language. Functions are listed under the six broad categories: imparting and seeking factual information; expressing and finding out attitudes; getting things done; socializing; structuring discourse; communication repair. Each can be broken down further. For example, 'imparting and seeking factual information' includes 'identifying', 'reporting' and 'correcting', and some sample utterances such as 'he is the owner of the restaurant' are given as exponents of 'identifying'.

In a similar way, eight general notions - existential, spatial, temporal, quantitative, qualitative, mental, relational and deixis - are listed. Existential is broken down into sub-categories such as presence / absence, availability / non-availability, and some exponents are given. A long list of specific notions for Threshold Level is also given, with exponents grouped together in fourteen themes: personal identification; house and home, environment; daily life; free time, entertainment; travel; relations with other people; health and bodycare; education; shopping; food and drink; services; places; language; weather.

A grammatical summary and word list are included in Threshold Level 1990, but it is emphasized that 'this apparatus of sentence formation, the grammar and the lexicon is not an end in itself, it is simply a tool for the performance of the communicative functions, which are what really matter.' A final aspect of Threshold which should be mentioned is its flexibility; the exponents of functions and notions which are listed are examples which could be replaced by others. Thus the specification could be adapted, by the inclusion of different vocabulary, for use by one particular professional grouping.

The flexibility of Threshold is also a feature of the way in which it has lent itself to the production of versions in other European languages. These are not translations of the original English language version, but independently created interpretations of the basic concept, which take account of differences in culture, and may vary considerably in the categories and exponents they include. At the time of publication of Threshold Level 1990, ten versions in languages other than English had appeared, and more have since been produced. All are available from the Council of Europe, Strasbourg, and from its regional agents in Europe and beyond.

Threshold, and the ideas which gave rise to it, influenced classroom teaching by leading to a movement away from teaching language structure by structure, towards situation and task-based learning, with an emphasis on materials which were either authentic or simulated authentic language use. Tasks which involved an 'information gap' were devised, in order to give learners a real need to communicate with each other. In testing, this also led to a greater use of authentic (or semi-authentic or simulated) materials, and away from discrete point testing towards test tasks which provide a context for language use.

With an emphasis on language as it is used in communication comes a need to analyse exactly which skills and abilities act together to give an individual

his level of competence in communication in a given language, whether it is his mother tongue or a second language acquired later in life. While language testing in the psychometric-structuralist era (Spolsky, 1975) paid little attention to defining the dimensions of language proficiency and communicative competence, developments in language testing since the mid-seventies have concentrated on this area.

Models of communicative competence

Models of communicative competence have provided an important focus since the late 70s as the basis of accepted practice in teaching and testing. Cziko (1982) has made a useful distinction with regard to the research in language testing during this era. He divides the research into two main categories, which he calls descriptive and working models of communicative competence. Descriptive models are ones which attempt to describe:

> all the components of knowledge and skills that a person needs to communicate effectively and appropriately in a given language.

Working models are defined as attempts to:

> show how components of communicative competence are interrelated psychologically to form a set of independent factors.

Descriptive models are illustrated by the work of Canale, Swain and Cummins, while working models have been researched by Oller, Palmer, Bachman and others.

The Canale and Swain model

Perhaps the best known descriptive model of communicative competence put forward in the eighties is the one proposed by Canale and Swain (1981; 1983) According to them, communicative competence encompasses four components - grammatical, sociolinguistic, discourse and strategic. Grammatical competence is concerned with the mastery of vocabulary, and the rules of word formation, sentential grammar, linguistic semantics, pronunciation and spelling. Sociolinguistic competence contributes to the individual's ability to communicate appropriately. It is the extent to which utterances are produced and understood appropriately in different sociolinguistic settings depending on factors such as the purpose of the interaction, the status of the participants and so on. It involves an awareness of the dos and don'ts of social interaction that are culture specific. Discourse competence refers to the mastery of the ways in which grammatical forms and meanings combine to achieve unified spoken or written texts. Finally, strategic competence refers to the mastery of verbal and nonverbal communication strategies to compensate for breakdowns in communication. Such strategies might include things like repetition, paraphrase and slower speech. Strategic competence is different from the other competencies postulated by Canale and Swain in that it interacts freely with the others. This model of communicative language ability considerably broadened the

perspectives of language testers in the eighties because it provided a framework for description, and thus validation, which had not been available before this time.

Cummins' views of communicative competence

Another descriptive model of communicative competence with an influence on the design of tests and interpretation of results was developed by Cummins (1979; 1983). His first model of communicative competence drew a distinction between cognitive/academic language proficiency (CALP) and basic interpersonal communication skills (BICS). While everybody is said to possess BICS the same is not true of CALP, which is strongly related to literacy skills. BICS is thus a kind of minimum competence, while CALP is acquired through education. This is why, according to Cummins, it takes language minority students much longer to attain grade/age appropriate levels in English academic skills than it does in face-to-face communication.

Cummins (1983) developed his position when he postulated that language proficiency can be conceptualized along two continua:

> First is a continuum relating to the range of contextual support available for expressing or receiving messages. The extremes of this continuum are described in terms of "context-embedded" versus "context-reduced" communication. They are distinguished by the fact that in context-embedded communication the participants can actively negotiate meaning (e.g. by providing feedback that the message has not been understood) ... context reduced communication, on the other hand, relies primarily (or at the extreme of the continuum exclusively) on linguistic cues to meaning and may in some cases involve suspending knowledge of the "real" world in order to interpret (or manipulate) the logic of the communication appropriately.

Cummins claims that interpersonal communication is normally context-embedded while context-reduced communication occurs in situations where linguistic precision is of great importance. Of course, the extent to which something is context-embedded or context-reduced is dependent to a large extent on the individuals concerned in the communicative event. However, the implication is that the less context there is the greater the effort to communicate will have to be. This position has implications for the types of task that one might include in a test in relation to the amount of context that the individual brings to the test, and encourages the view that tests can never be fair to everybody. In practical terms this is something that test designers have to accept but it is an important point to bear in mind in the selection of test materials and tasks. It supports the commonly held view that if test items do not provide a familiar context for test takers, then the items may be more difficult.

The second continuum in the Cummins model relates to the amount of cognitive involvement in a task or activity. Cummins defines cognitive involvement as:

the amount of information that must be processed simultaneously or in close succession by the individual in order to carry out the activity.

Language tasks are categorized in relation to the amount of context supplied and the degree to which they are cognitively demanding. However, it is difficult to define the notion of context very precisely because it will be different from one individual to the next. Furthermore, what might be cognitively demanding at the beginning of a learning cycle may not be at a later stage.

One of the most important implications of this model for language testing is that it encourages the test designer to focus more on the learner as an individual. In order to use the contextual and cognitive dimensions it is inevitable that the test designer will consider the background characteristics of the people being tested.

Morrow and authentic communication

Morrow (1979) does not really attempt a definition of communicative competence, but his ideas are important and relevant to this area of study, particularly as he lists some of the features of authentic communication that should be taken into account in communicative language test design if tests are to be valid:

i Communication is interaction based, in that what is said or written by an individual depends crucially on what is said or written to him/her.

ii Communication is unpredictable and data has to be processed in real time.

iii Communication requires a context which is situational as well as linguistic.

iv Communication is purposeful in that an individual must be able to recognize why utterances are addressed to him/her and produce relevant responses that will achieve the desired purpose.

v Communication requires performance, that is, the ability to use language in real situations.

vi Communication involves the use of authentic texts.

vii Communication is behaviour-based in that it has an outcome.

Drawbacks of Morrow's approach

These features are explained in greater detail in Morrow's article Communicative language testing: Revolution or evolution (1979). Although this is a useful list and of value in communicative test design, it has been criticized (Alderson, 1981; Weir, 1981; Moller, 1981) because Morrow does not define terms like communicative proficiency, language competence, performance test and behavioural outcome. Nor does he explain adequately

how the seven features outlined above are to be taken into account in the design of communicative language tests and how they should be measured or weighted. However, since the late seventies, this approach has encouraged the production of tests that look more appealing and realistic with strong face validity.

Bachman and communicative language ability

By the late 1980s a series of developments were taking place in language testing. Bachman (1990) presented his first comprehensive view of CLA (communicative language ability), which clearly grew out of the work of Canale and Swain. He suggested that CLA consists of knowledge or competence plus the ability to execute that competence in appropriate language use. It is broken down into language competence, strategic competence and psychophysiological mechanisms. These interact with the language use context and the user's knowledge structures.

Language competence is broken down into organizational competence and pragmatic competence. Organizational competence is further broken down into grammatical competence (vocabulary, morphology, syntax and phonology/graphology) and textual competence (cohesion and the rhetorical organization of written or spoken discourse). Pragmatic competence concerns the relationship between the linguistic signals given in communication and both the language users and the context of communication. It is broken down into illocutionary competence and sociolinguistic competence. Illocutionary competence has to do with the user's ability to employ language functions: ideational (used to express ideas, knowledge, feelings), manipulative (used to affect the world around the user), heuristic (used to extend knowledge of the world) and imaginative (used for creative purposes - jokes, poetry, etc.) Sociolinguistic competence has to do with the appropriate use of language in context. It consists of sensitivity to differences in dialect or variety of language, register, naturalness (what a native speaker would use) and ability to interpret cultural references and figures of speech.

In this version of Bachman's model, language competence interacts with strategic competence, made up of the ability to assess, plan and execute appropriate interactional language use by the most effective means, and psychophysiological competence, which includes the anatomical means of speech.

In the models of language ability outlined above, one feature of particular interest is the varying view expressed of the nature of strategic competence. It is clear that all writers on this subject regard it as an important area of competence, but it is perhaps the most difficult to define.

The term 'strategic competence' has been used in some quite different ways. As mentioned above, Canale and Swain consider it to be the ability to use both verbal and nonverbal means of repairing breakdowns in communication. If, for example, it is clear that what a speaker has said has not been understood, he may try to compensate for this by repeating the same words

more slowly and clearly, or by trying to convey the same information using different vocabulary. In Threshold Level the same idea seems to be expressed in a section entitled Compensation Strategies, which describes how the learner has to be able to develop strategies for coping adequately with 'unpredicted demands as well as with failures of recall.' Bachman, on the other hand, takes what appears to be a more positive, less defensive view of strategic competence, seeing it as the ability to assess and plan in order to execute an appropriate form of language use.

It is possible to take a view of strategic competence which embraces both assessment and planning and strategies for compensation and repair. It follows from this that strategic competence may operate, depending on the circumstances, in ways of which the language user is consciously aware, or at a subconscious level. The effort involved in learning a second language, and the learner's inevitable lack of native speaker competence mean that strategic competence is likely to be closer to the area of conscious awareness for the language learner than it is for the native speaker.

In turning to the implications of this for language teaching and testing, one can distinguish between two kinds of tasks. Because they demand an immediate response, oral tasks make it necessary for the learner to acquire some forms of routinized communication, and to develop ways of dealing with breakdowns in communication. Written tasks, because they do not demand an immediate response, may include, or even demand, conscious planning. Self-assessment and the learner's own development of control over the learning process are closely tied into the growth of strategic competence. As far as testing is concerned, it may be possible to structure tasks set in a test of speaking in such as way as to make it likely that the candidate will have to resort to compensation strategies. The examiner would take into account how well such strategies were used. However, there is a problem here, in that the most successful compensatory strategies are self-concealing! It would be easier to test the planning element in strategic competence by building a need for planning into written test tasks, in such as way that the planning can be assessed as a separate element from task performance.

Modifications to the Bachman model

Bachman and Palmer (1996), present a modified version of the language ability model. The first characteristic of the model to consider is language knowledge. This includes all the features that one would associate with systemic knowledge of a language in conjunction with the characteristics of usage.

Figure 1

. Language Knowledge

Organisational Knowledge		Pragmatic Knowledge		
Grammatical Knowledge	Textual Knowledge	Lexical Knowledge	Functional Knowledge	Sociolinguistic Knowledge
Syntax Morphology Graphology/ Phonology	Rhetorical organisation Cohesion	Semantic properties Denotation Connotation	Ideational Manipulative Heuristic Imaginative	Conventions of language use Dialect/Variety Register Naturalness

(Bachman & Palmer 1996)

But whilst language knowledge is obviously an essential prerequisite to using language, its successful application is crucially affected by its interaction with other mental processes namely, knowledge schemata and affective schemata. Knowledge schemata refers to knowledge and experience of the world whereas affective schemata refers to emotional memories. The model proposes that these three mental characteristics are activated and utilized by a metacognitive process involving three phases; assessment strategies, planning strategies and goal-setting strategies.

In a conversation, for example with a friend, a speaker may make an assessment of the appropriate topic based on his knowledge schemata, from which he assesses and plans the language at his disposal for the communicative goal he has set himself. But after initiating the conversation if new information is discovered such as the friend's new promotion, then this will precipitate further assessment, planning and goal-setting and use of alternative language. The model is clearly a dynamic one which, on the one hand, involves continual interaction between the metacognitive processes and the user's language knowledge, affective schemata and knowledge schemata and, on the other hand, involves interaction between the language user and the language use context.

A model of language ability is of use to the language tester because it provides a 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, and makes possible the devising of such useful tools 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. A test which can do this satisfactorily is a **valid** test. The exact nature of validity and how it can be achieved has been, and still is, a subject of debate in language testing.

Validity

The Condition of Validity

Traditionally the concept of validity is generally defined as the extent to which a test measures what it is supposed to measure (Pratt, 1980; Popham, 1981; Priestly, 1982; Carroll and Hall, 1985), or the extent to which it provides information relevant to the decision that is to be made on the basis of the test results (Thorndike and Hagen, 1977). The attempt to establish validity can be made in a number of ways: by comparing test content with the syllabus on which it was based; by comparing the results of a test purportedly measuring a particular skill or trait with another already established test measuring the same trait; by comparing test performance with eventual success in a given area; and by attempting to establish whether there are grounds to claim that the skills being tested do indeed reflect the competence theory that underlies a given test.

Aspects of validity and how they can be established

Many different ways of establishing validity have been put forward. For some years discussion of this question focused on defining several different aspects of validity. Standards for Educational and Psychological Tests (1974), with which by Popham (1981), Cronbach (1970), and Thorndike and Hagen (1977) are in agreement, suggests three basic types of validity: content, criterion-related and construct validity. In addition to these, a definition of a fourth type, face validity, has been put forward.

Content Validity

Content validity, sometimes referred to as the principle of “inclusiveness” (McCormick and James, 1983) concerns the extent to which a test covers the content that it is supposed to. This aspect of validity is very important in the context of achievement and progress tests at the classroom and school level (Deale, 1975). If we are considering an achievement test, the content to be covered will be found in the syllabus and course books. Content validity can be established by comparing what is in the syllabus, for example, and what is in the test. In theory, the closer the match, the better the content validity. If proficiency is being tested, then it tends to be the test constructor who defines the content of the test. Moller (1982, p.37) gives some thought to this issue:

> “Content validity, together with reliability, will ensure that a test adequately reflects the objectives and linguistic content laid down in the syllabus. In the case of a proficiency test, however, the test constructors themselves decide the `syllabus’ and the universe of discourse to be sampled. The sampling becomes less satisfactory because of the extent and indeterminate nature of that universe. Thus the evaluator looking for content validity is really assessing the test constructors’ definition of proficiency.”

In fact, the same argument can be used with regard to achievement testing. Somebody, somewhere, generally designs a syllabus and, on the basis of this, materials writers prepare textbooks. The appropriacy of the content of the syllabus, and the extent to which the materials then reflect it are subjective decisions for the most part. Thus, a published language course may be based on the Council of Europe's Threshold Level. As a course it may or may not adequately reflect Threshold. Institutions will then use the course and may even write tests to accompany it. These tests may be valid in the context of the materials, but there is no guarantee that the materials are actually valid in the context of Threshold - nor indeed that Threshold itself is valid.

Criterion-related Validity

Criterion-related validity refers to the extent to which a test is predictive and/or concurrent i.e. measuring the same thing as another, already proven test. The most usual way of establishing criterion-related validity is by correlation.

To establish the concurrent validity of a test it is normal that students' results on another test purportedly measuring the same trait as the test under investigation are correlated with that test. The problem arises in that it is difficult to show conclusively that two tests are indeed measuring the same things in the same way. This has long been perceived as a difficulty but recent research into trait and method effect (Bachman and Palmer, 1983; Shohamy, 1984) has made it very clear that there can be no guarantees that methods are equivalent. Thus, when new item types are developed, it is very questionable that traditional concurrent validation procedures are applicable.

It is sometimes the case that concurrent validity can be demonstrated by correlating test results with either teacher ratings or grades (Chaplen, 1970 ;). Ingram (1974) goes as far as to say that teachers' ratings are the best method of establishing the concurrent validity of a test. However, defining criteria for teachers to use is generally a problematic issue (Moller, 1982, p.52), as is the standardization of teachers' views of what those criteria mean.

If establishing concurrent validity is fraught with difficulties then so is determining predictive validity. Many studies have focused on the academic context since this is where most students take proficiency tests, and where most data are available for investigation. Moller (1982) reviews a number of predictive validity studies and makes the point that most of them have used non-linguistic criteria, such as Grade Point Averages in their work. Numerous factors can affect the results of a predictive validation study such as the amount of exposure that subjects get to the target language, their willingness and ability to improve, the extent to which language issues are focused on by their tutors and so on. In addition, in many predictive validity studies the samples used are unavoidably biased in the sense that studies can only be carried out on subjects who achieve the sort of score that allows them to go to a university in an English speaking country. All those who do not achieve the appropriate grade are excluded. Whether some of them would have succeeded is open to question.

Construct Validity

To satisfy the condition of construct validity, a test must be shown to measure the psychological constructs that it is hypothesized to be testing. It is a basic assumption that tests are intended to provide us with information on some real world phenomenon, characteristic or behaviour. They are generally an indirect or operational way of attempting to describe the extent to which individuals possess a theoretically postulated characteristic or construct.

The process of construct validation, according to Walsh and Betz (1985) may be broken down into three stages:

> "First, the construct of interest is carefully defined and the hypotheses regarding the nature and extent of its relationships to other variables are postulated. Second, an instrument designed to measure that construct is developed. Third, after the degree to which the test is reliable has been examined, studies examining the relationship of the test to other variables (as formulated in the hypotheses about the construct of interest) are undertaken."

Numerous statistical techniques are available to investigate construct validity, although one of the most popular ones in recent years has been the factor analysis of the intercorrelations of subtests or items in order to establish how many dimensions or traits may be required to summarize or explain test performance. For example, if a test is designed to measure more than one trait and yet a factor analysis yields a single "general" factor then, from this point of view at least, the test could not be said to possess construct validity.

In language testing research construct validation attracted relatively little interest until the late seventies when suddenly a spate of research, mostly in the United States, began to focus on it. Most of this research can be viewed, according to Weir (1984, p.65):

> "... principally as the a posteriori statistical validation of whether a test has measured a construct which has a reality independent of other constructs. The concern is much more with the a posteriori relationship between a test and the psychological abilities, traits, constructs, it has measured than with what it is that it should have elicited in the first place."

As important as the a posteriori validation of a test is the a priori establishment of the appropriateness of test content. That is to say, there needs to be a clear definition of what is to be tested before we can attempt to establish that it has been tested. Ebel (1983) points out in fact, that a major reason for problems of test validation is an overemphasis on the need for empirical validity data, and a failure to recognize the primary importance of explicit verbal definitions of what the test is intended to measure.

A greater focus on a priori validation inevitably leads to an overlap between content and construct validity.

Face Validity

Although the concept of face validity is an important one, it is not generally included as one of the three major types of validity due to the questionable role that it plays in the validation process.

Face validity, the extent to which a test looks as if it is testing the right thing in the eyes of students, teachers, and sponsors has also received much attention in discussions about the validity of language tests. There is no generally accepted procedure for determining whether a test has face validity, and this has led some to argue that it should have no place in the discussion of test validity (Bachman et al., 1981).

Stevenson (1985), while supporting the movement in language testing towards more performance based formats, warns that:

> "Face validity is the mere appearance of validity to the metrically-naive observer. It provides the psychometrically unsophisticated self-assurance that allows someone simply to look at a test and, without further technical examination, conclude: "I know a valid test when I see one."

Stanley and Hopkins (1972, p. 105) also point out that face validity is very much on a naive and superficial level and that it is dangerous to attribute too much importance to it. While a test with good content validity will generally have face validity, the reverse is much less likely to be true.

Bachman (1990, pp. 285-289) reviews discussions in the literature dating from 1947 (Mosier). He notes that most comments on the notion of face validity (or 'test appeal' as he calls it) have been negative and he proposes a 'post mortem' on the term. He claims that the 'final interment' was 'marked by its total absence from the most recent (1985) edition of the ''Standards'' ' (APA). For Bachman '...the ''bottom line'' in any language testing situation, in a very practical sense, is whether test takers will take the test seriously enough to try their best, and whether test users will accept the test and find it useful. For these reasons, test appearance is a very important consideration in test use.' (1990, p.288).

In summary, therefore, there can be no doubt that a test should look as if it is testing the right things in the right ways, whatever that may mean. However, there can be equally little doubt that it is inappropriate to approach test validation primarily from the angle of face validity, which should naturally arise out of a real concern to ensure that the content of a test is valid.

Recent Views Of Test Validity

More recently, there has been a movement away from regarding validity as something which can be broken down into different aspects and labelled as such, and towards regarding it instead as an ongoing and integrated process of validation. Views began to change in the early 80s, when Cronbach, for example, expressed the view that "all validation is one" (1980, p.99) and his view reflected the emphasis that began to be placed on *construct validity* as the essential component of the unitary conception.

In current thinking, therefore, validity is considered a unitary concept. Instead of referring to *types* of validity, texts now refer to categories of *validity evidence* (see APA Standards 1985). Bachman (1990), following this line of thought, defines validation as 'collecting evidence of validity'. The process of validation can therefore be considered a form of scientific enquiry which

requires systematic collection of information (i.e. evidence of test validity), which in turn allows for the appropriate interpretation of results by test users (i.e. consequences of test use).

Messick (1989) has been influential in promoting this view of validity in his writings. He has argued that, although construct validity provides crucial evidence for the interpretation of results, it is not enough on its own. He proposes *a progressive matrix* which has construct validity as a key component of each cell but which also takes into account the justification for testing and the outcomes which follow from the adoption of a particular test or testing system:

Messick's Matrix (1989, p. 20)

	Function of outcome	
Source of Justification	*Test interpretation*	*Test use*
Evidential basis	Construct validity	Construct validity + Relevance/utility
Consequential basis	Value implications	Social consequences

According to Messick's view, construct validity should be considered as just one justification (along with relevance, utility, value implications and social consequences) for *the use and interpretation of tests*. Content-related validity, for example, is thus viewed as a variety of evidence which supports construct validity and provides necessary information for the interpretation of test results.

Messick's framework has yet to be applied to an operational test (partial applications have been published) due in no small part to the difficulty in fully understanding how Messick himself saw this happening. However, as we will see in the following section, the main ideas that set Messick's conceptualisation of validity above the rest (the 'unitary' concept, that validity is a single entity, for which the developer should provide a range of evidence; and the notion of consequential validity, where the developer is responsible for and is expected to take into account the intended and untended impact of a test) have been very much the driving force behind the latest trend in test validation.

Operational Frameworks for Test Validation

The most recent trend in validity theory is to try to step beyond the purely theoretical – and often not clearly practical – modelling of validity to a more practical and operational description of validity. These descriptions typically take the form of a framework which is seen by its supporters as providing a theoretically sound rationale both to test development and validation.

The most influential of these frameworks have been presented by Mislevy (in the USA) and Weir (in Europe).

Evidence Centred Design (ECD): Mislevy

In a series of papers in the late 1990s and early 2000s, Bob Mislevy and his associates presented a framework for test. The basis for this framework is the focus on a series of four stages:

Stage	Title	Brief Description
1	Domain Analysis	A broad-ranging area which involves the collection of information from a range of sources related to the domain. This broadly relates to the area of needs analysis.
2	Domain Modelling	This relates to the broad descriptions of what Mislevy (2003: 6) calls the 'evidentiary relationships' between the key paradigms that define a domain (claims, evidence and task paradigms).
3	Conceptual Assessment Framework (CAF)	This relates to the specification of the operational test or assessment, from a range of perspectives. *Student Models:* test taker characteristics relevant to the assessment *Task Models:* language elicitation devices *Evidence Models:* task and test scoring *Assembly Model:* how the test is assembled from chosen tasks *Presentation Model:* how the tasks should be presented, the interactions managed and the performances recorded.
4	Operational Assessment	The operational stage of the test development cycle.

As can be seen in the above table, the *Domain Analysis* stage is the preliminary, data gathering, phase of development. Here the test developer attempts to describe the knowledge and tasks that define the domain from a range of sources.

Following on from this stage is *Domain Modelling,* seen by McNamara (2003: 9) as 'the most crucial stage of the process'. The three sub-stages are: claims, evidence and tasks. The idea here is that the construct is defined by first clarifying the claims we hope to make concerning the candidates and then indicating the kind of evidence required to substantiate these claims. Finally, the developer must decide (in the *Task* sub-stage) how the evidence should be gathered, i.e. identifying the most efficient techniques for observing the skills/abilities being assessed.

Domain Modelling (Substantive Argument)		
CLAIMS	**EVIDENCE**	**TASKS**
The characteristics of students and aspects of proficiency they reflect.	The characteristics of what students say and do – what would a student need to say or do to establish a basis for the claim about them? (How to identify evidence in student work – criteria)	Kinds of situations that might make it possible to obtain this evidence, which minimise construct-irrelevant variance.

Source: McNamara 2003: 9

In the *CAF* stage, the developer creates the detailed technical specification or blueprint from which the final assessment will emerge. This document should be explicitly based on the earlier stages of development in order to ensure that the operational assessment reflects the construct definition (claims and evidence) and solution (tasks) described in the *Domain Modelling* stage. The final stage is self explanatory. The result of all this thinking, planning and describing is an operational assessment.

This framework has been used as part of the theoretical foundation of the latest version of the TOEFL test from ETS in the USA. However, the extent to which Mislevy's ideas will be put into practice in the general domain of language testing is unclear. The ideas are complex and appear more suited to large scale assessment systems that rely heavily on technology for their delivery, analysis and reporting mechanisms. Nevertheless, the framework offers an interesting perspective on the linking of the theoretical underpinnings of an assessment to its operational description and specification. This suggests that it would be of interest both to the test developer and to the critical reviewer.

At the same time as Mislevy and his associates were putting together their framework, Cyril Weir and his colleague Barry O'Sullivan at the Centre for Research in Testing, Evaluation and Curriculum in ELT (CRTEC) in London, were developing their Test Validation Frameworks, again with an evidence-based approach, though here with more explicit reference to the socio-cognitive approach to language use.

Test Validation Frameworks (Weir)

Weir (2005) proposes a comprehensive framework which, he argues, can form the basis of any test development and validation project. In the outline presented below, we can see that there are a number of elements, each of which should be attended to by the test developer. Weir (2005:48) argues that test developers are obliged to seek to address all of the following questions:

- How are the physical/physiological, psychological and experiential characteristics of candidates catered for by this test? (*Test-taker*)

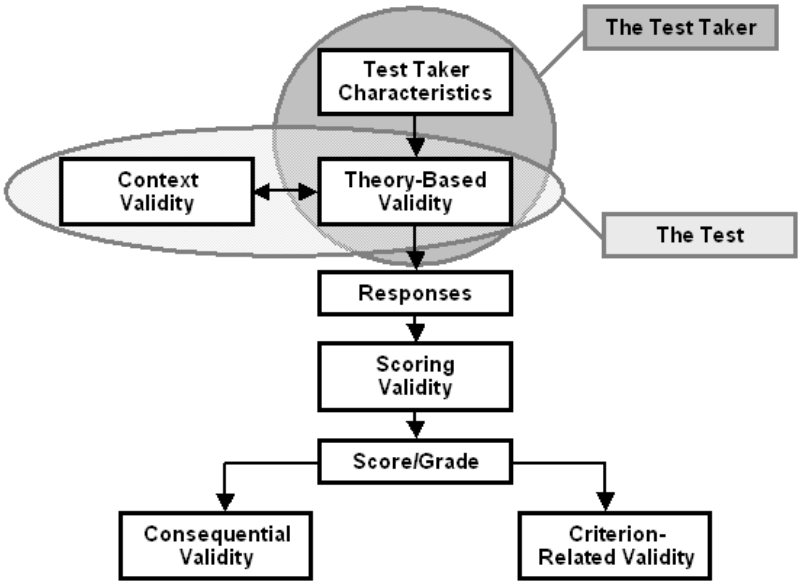
- Are the characteristics of the test task(s) and the administration fair to the candidates who are taking the test? (*Context validity*)

- Are the cognitive processes required to complete the tasks appropriate? (*Theory-based validity*)

- How far can we depend on the scores on the test? (*Scoring validity*)

- What effects does the test have on its various stakeholders? (*Consequential validity*)

- What external evidence is there outside of the test scores themselves that the test is doing a good job? (*Criterion-related validity*)

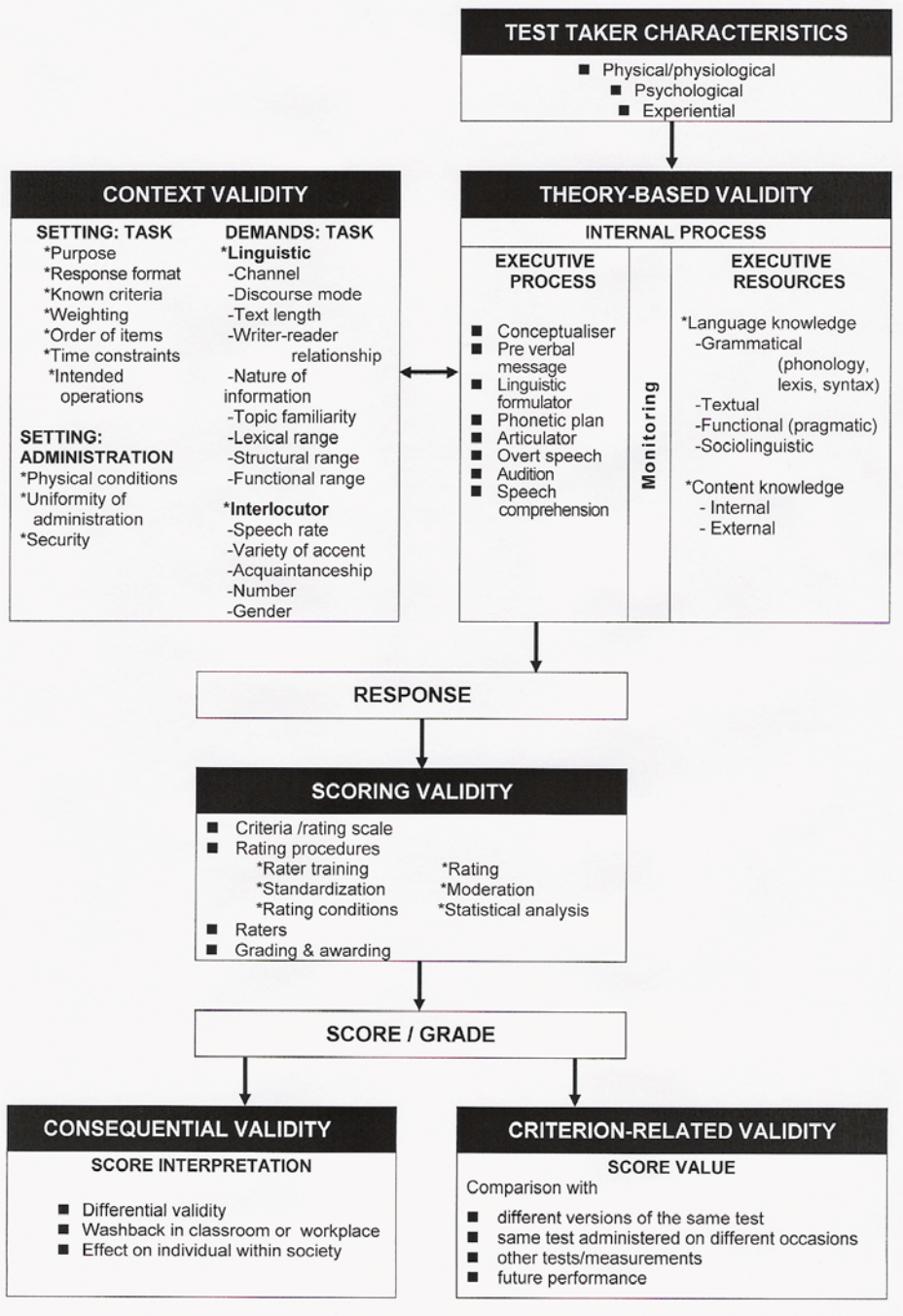


For descriptive purposes the elements of the model are presented as being independent of each other. However, Weir argues that there is a 'symbiotic' relationship between context validity, theory-based validity and scoring validity, which he sees as constituting *construct* validity. Examples of this symbiotic relationship include

- decisions taken with regard to parameters in terms of task context will impact on the processing that takes place in task completion

- making scoring criteria are made known to candidates in advance will affect executive processing in task planning and completion

Weir proposes four frameworks, one for each of the four skills. In order to demonstrate what the frameworks look like and how they can be applied, one is shown as an example, that for speaking.



The various parameters are briefly described in the following tables

TEST TAKER CHARACTERISTICS	
Physical/Physiological	
Short term ailments	*Toothache, cold etc.*
Longer term disabilities	*Speaking, hearing, vision (e.g., dyslexia)*
Age	Suitability of materials, topics etc. Demands of tasks (time, cognitive load etc)
Sex	Suitability of materials, topics etc.

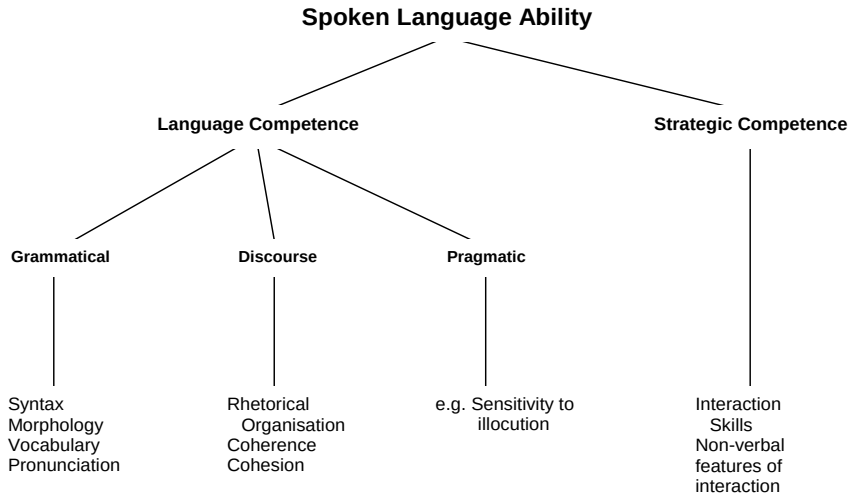
Psychological	
Memory	Related to task design, also to physical characteristics
Personality	Related in speaking primarily to task format (e.g. number of participants in an event – solo, pair, group, etc.)
Cognitive Style	e.g. preferred learning style may impact on task performance
Affective Schemata	How the candidate reacts to a task. Can be addressed by the developer through carefully controlled task purpose
Concentration	Related to age and also (particularly in listening and reading) to length and amount of input
Motivation	Among other things this can be related to task topic or to task/test purpose
Emotional state	An example of an unpredictable variable. Difficult to deal with, though may be approached from the same perspective as Motivation or Affective Schemata.
Experiential	
Education	This can be formal or informal and may have taken place in a context where the target language was either the principal or secondary language
Examination Preparedness	Can relate either to a course of study designed for this specific examination, examinations of similar design or importance or to examinations in general.
Examination Experience	Again can relate to this specific examination, examinations of similar design or importance or to examinations in general.
Communication Experience	Can relate to any of the above, e.g. where communication experience is based only in classroom interactions or where the candidate has lived for some time in the target language community and engaged in 'real' communication in that language.
TL-Country Residence	Can relate to Education (i.e. place of education) or to Communication Experience (e.g. as a foreign or second language)

THEORY BASED VALIDITY	
INTERNAL PROCESSES **(based on Levelt 1989)**	
Conceptualiser	conceiving an intention, selecting relevant information to be expressed to realize this purpose, ordering information for expression, keeping track of what was said before; paying constant attention to what is heard and own production, drawing on procedural and declarative knowledge. Speaker will monitor message before they are sent into the formulator.
Pre verbal message	product of the conceptualization
Linguistic formulator	includes grammatical encoding and phonological encoding which accesses lexical form
Phonetic plan	an internal representation of how the planned utterance should be articulated; internal speech
Articulator	the execution of the phonetic plan by the musculature of the respiratory, the laryngeal and the supralaryngeal systems

Overt speech	
Audition	understand what is being said by others or self, i.e. interpret speech sounds as meaningful words and sentences
Speech comprehension	access to various executive resources e.g. lexicon, syntactic parser, background knowledge. A representation is formed of the speech in terms of its phonological, morphological, syntactic and semantic composition. Applies to both internal and external overt speech.
MONITORING	both of internal and external speech can be constantly in operation though sometimes this filter is switched off. The system through which internal resources are tapped in response to demands of executive processing.
EXECUTIVE RESOURCES	
Content knowledge	
Internal	The test-taker's prior knowledge of topical or cultural content (background knowledge)
External	Knowledge provided in the task
Language knowledge (all references to Buck 2001)	
Grammatical	literal semantic level: includes phonology, stress, intonation, spoken vocabulary, spoken syntax
Discoursal	longer utterances or interactive discourse between two or more speakers: includes knowledge of discourse features (cohesion foregrounding, rhetorical schemata and story grammars) and knowledge of the structure of unplanned discourse
Functional	function or illocutionary force of an utterance or longer text + interpreting the intended meaning: includes understanding whether utterances are intended to convey ideas, manipulate, learn or are for creative expression, as well as understanding indirect speech acts and pragmatic implications
Sociolinguistic	the language of particular socio-cultural settings + interpreting utterances in terms of the context of situation: includes knowledge of appropriate linguistic forms and conventions characteristic of particular sociolinguistic groups, and the implications of their use, or non-use, such as slang, idiomatic expressions, dialects, cultural references, figures of speech, levels of formality and registers

In terms of Executive processes, it is important to realise that the test developer can only realistically hope to directly impact on the first stage of the process – for example by making clear to the candidate the criteria which will be used in assessing performance, by ensuring that the rubrics for the task are unambiguous, etc. Research can show that the cognitive and meta-cognitive strategies used by the candidate in monitoring the interaction between the executive resources and processes reflect those used in the real world domain the test is attempting to replicate. The executive resources available to the candidate include *Content Knowledge* (either brought to the test event through existing background knowledge or learnt during the event from information provided in the test) and *Language Knowledge* (expressed here through the most commonly quoted model – that of Bachman 1990, Bachman & Palmer 1996). This latter type of knowledge is usually described in terms of the theoretical model of language competence that drives the test,

see for example the model presented by Saville and Hargreaves (1999: 45) upon which the Cambridge ESOL examinations are based (below).



CONTEXT VALIDITY	
Settings: Task	
Purpose	The requirements of the task. Allow candidates to choose the most appropriate strategies and determine what information they are to target in the text in comprehension activities and to activate in productive tasks. Facilitates ***goal setting*** and ***monitoring.***
Response format	How candidates are expected to respond to the task (e.g. MCQ as opposed to short answers). Different formats can impact on performance.
Known criteria	Letting candidates know how their performance will be assessed. Means informing them about rating criteria beforehand (e.g. rating scale available on WEB page).
Weighting	Goal setting can be affected if candidates are informed of differential weighting of tasks before test performance begins.
Order of Items	Usually in speaking tests this is set, not so in writing tests.
Time constraints	This can relate either to pre-performance (e.g. planning time), or during performance (e.g. response time)
Intended operations	A broad outline of the language operations required in responding to the task. May be seen as redundant as a detailed list is required in the following section.
Demands: Task *[note: this relates to the language of the INPUT and of the EXPECTED OUTPUT]*	
Channel	In terms of input this can be written, visual (photo, artwork, etc), graphical (charts, tables, etc.) or aural (input from examiner, recorded medium, etc). Output depends on the ability being tested.
Discourse Mode	Includes the categories of genre, rhetorical task and patterns of exposition

Text Length	Amount of input/output
Writer/speaker relationship	Setting up different relationships can impact on performance (e.g. responding to known superior such as a boss will not result in the same language as when responding to a peer).
Nature of Information	The degree of abstractness. Research suggests that more concrete topics/inputs are less difficult to respond to that more abstract ones.
Topic familiarity	Greater topic familiarity tends to result in superior performance. This is an issue in the testing of all sub-skills
Linguistic	
Lexical Range	these relate to the language of the input (usually expected to be set at a level below that of the expected output) and to the language of the expected output. Described in terms of a curriculum document or a language framework such as the CEFR.
Structural Range	
Functional Range	
Interlocutor	
Speech Rate	Output expected to reflect that of L1 norms. Input may be adjusted depending on level of candidature. However, there is a danger of distorting the natural rhythm of the language, and thus introducing a significant source construct-irrelevant variance.
Variety of Accent	Can be dictated by the construct definition (e.g. where a range of accent types is described) and/or by the context (e.g. where a particular variety is dominant in a teaching situation).
Acquaintanceship	There is evidence that performance improves when candidates interact with a friend (though this may be culturally based).
Number	Related to candidate characteristics – evidence that candidates with different personality profiles will perform differently when interacting with different numbers of people.
Gender	Evidence that candidates tend to perform better when interviewed by a woman (again can be culturally based), and that the gender of one's interlocutor in general can impact on performance.

SCORING VALIDITY	
Criteria/Rating Scale	The criteria must be based on the theory of language (Language Knowledge) outlined in the Theory Based Validity section and reflected again in the Demands: Task section of Context Validity. They should also reflect 'actual' language production for the task or tasks included in the examination.
Rating Procedures	
Training	There are a number of different approaches to training, and there is evidence that training improves harshness, consistency and ability to stay on standard.
Standardisation	As part of any training regime, rates must internalise the criterion level (e.g. pass/fail boundary) and this should be checked using a standardisation procedure (or test if you like).
Conditions	Attempts should be made to ensure that all rating/examining takes place under optimal conditions. Where possible, these conditions should be set, so that all examiners have an equal opportunity to perform t their best.

Moderation	This involved monitoring the performance of raters to ensure that they stay on level.
Analysis	Statistical analysis of all rater performances will ensure that individual candidates will not loose out in situations where examiners are either too harsh/lenient or are not behaving in a consistent manner.
Raters	When we discuss the candidate (in terms of physical, psychological and experiential characteristics) we should also consider what we know of the examiners in terms of these same characteristics. Little research has been undertaken in which these have been systematically explored from the perspective of the rater.
Grading & Awarding	The systems that describe how the final grades are estimated and reported should be made as explicit as possible to ensure fairness.

Since the notion of Criterion-Related Validity as seen by Weir is the same as the traditional view (see the relevant section above) this element of the framework will not be discussed at this point, but the reader will be referred to the appropriate section above.

The final element of the framework relates to Consequential Validity. While it is not completely clear how this element of the framework 'fits' into an overall validity argument, the argument presented by Weir will be briefly outlined.

A personal perspective on consequential validity is that it does not exist as a separate element but instead should be seen as a more global aspect of test development which informs an ethical approach to all stages of development.

CONSEQUENTIAL VALIDITY	
Differential Validity	Refers of post-test analysis of response data to highlight potential instances of bias towards or against particular sub-sections of the test population.
Washback	The impact of the test on learning and teaching. Tests have been used in an attempt to influence teaching, though the evidence seems to suggest that at best only superficial changes occur unless teachers are made part of the innovation process. While there is a lot of evidence of the existence on washback, there is still no widely accepted description or definition of what exactly it is (though see the volume by Cheng, Watanabe & Curtis (2004) for a range of studies that have moved on our understanding of the area.
Impact on individual within society	The effect of a test on the wider community is possibly the most difficult area of all to investigate and the one most likely to be overlooked as it demands going beyond the immediate stakeholders in the testing process.

Weir's framework has been used to describe existing examinations (see Weir & Shaw and O'Sullivan & Green, both forthcoming), as a basis for demonstrating the specificity of language for specific purposes (see O'Sullivan 2005) and as the basis for the creation of detailed test

specifications (see QALSPELL 2005 and EXAVER 2005). It has also been adapted for use in examinations for other subjects (e.g. the Mathematics Placement Test in the United Arab Emirates). The influence of the framework, particularly in the European context is growing, at least partly due to its relative ease of application to a whole range of examination in a range of contexts.

EXERCISES

Materials for the Guidance of Test Item Writers

1. Look at the following set of quotations, and discuss or consider the following questions.

To what extent do you agree or disagree with each?
What are the implications for test design?

1.
The primary function of language is for interaction and communication.

(Richards and Rodgers, 1988: 70)

2.

The major problem in learning a foreign language is to master the structure of that language and this problem requires almost exclusive attention.

(Roberts, 1982: 99)

3.

Language is a set of habits. Foreign language learning is basically a process of mechanical habit formation.

(Richards and Rodgers, 1988: 51, on 'Audiolingualism')

4.

Communication involves freedom and unpredictability.

(Xiaoju, 1990: 61)

5.

Teaching materials should be as authentic as possible at all times.

(Wright, 1987: 76)

2. Take two examples of language tests you are familiar with.

Try to relate each one of them to one of the views on language or models of language ability outlined in this module.

APPENDIX A

Recommended reading

The following is a short list of some of the essential texts on which this module is based. A more extensive bibliography is also given in Appendix B.

Bachman L F and Palmer A S. 1996. *Language testing in practice*. Oxford: Oxford Univeristy Press

EXAVER 2005. Project Website. http://www.uv.mx/exaver/index.html. Website accessed April 2005.

Heaton, J.B. 1975. *Writing English Language Tests*. London: Longman.

Lado, R. 1961. *Language Testing*. London: Longman.

McNamara, T. 2003. *Validity and Reliability in the Senior School Curriculum: new takes on old questions*. Invited Presentation, Australian Curriculum, Assessment & Certification Authorities (ACACA), National Conference, Adelaide, July 31st 2003.

Mislevy, R. 2003. *On the Structure of Educational Assessments*. CSE Technical Report 597. University of California, Los Angeles: Centre for the Study of Evaluation.

Munby, J. 1978. *Communicative Syllabus Design*. Cambridge: Cambridge University Press.

O'Sullivan, B. 2005. *Issues in Business English Testing: the BEC revision project*. Cambridge: Cambridge University Press.

O'Sullivan, B. & Green, A. (forthcoming) *Examining Speaking*. Cambridge: Cambridge University Press
QALSPELL. 2005. *Project Handbook.*

van Ek, J.A. and J.L.M. Trim. 1991. *Threshold Level 1990*. Strasbourg: Council of Europe.

Weir, C. & Shaw, S. (forthcoming) *Examining Writing*. Cambridge: Cambridge

Weir, C. 2005. *Language Testing and Validation: an evidence-based approach*. Oxford: Palgrave.

Wilkins, D. 1976. *Notional Syllabuses*. Oxford: Oxford University Press.

APPENDIX B

BIBLIOGRAPHY

Alderson, J.C.1981. Reaction to the Morrow paper (3). In Alderson, J.C and A. Hughes (eds.). 1981. *Issues in Language Testing*. ELT Documents 111. London: The British Council.

Alderson, J.C. and A. Hughes. 1981. *Issues in Language Testing*. ELT Documents 111. London: The British Council.

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

Blalock, H.M. 1980. *Sociological Theory and Research*. NY: Macmillan.

Bolton, S. 1985. *Die Gütebestimmung kommunikativer Tests*. Tübingen

Brumfit, C. 1980. *Problems and Principles in English Language Teaching*. Oxford: Pergamon Press.

Canale, M. and M. Swain. 1981. A theoretical framework for communicative competence. In Palmer, A. S., Groot, P. G. and Trosper, S.A.(eds.) *The construct validation of tests of communicative competence.* Washington, D.C., TESOL, 31-36.

Carroll, J.B. 1968. 'The psychology of language testing' in Davies 1968: 46-69

Cronbach, L.J. 1990. *Essentials of Psychological Testing*. NY: Harper and Row.

Cummins, J. 1979. Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters. *Working Papers in Bilingualism*. 19: 97-205.
Finocchiaro, M. and C. Brumfit. 1983. *The Functional-Notional Approach: From Theory to Practice*. Oxford: Oxford University Press.

Davies, A. (ed.) 1968a. *Language Testing Symposium. A Psycholinguistic Perspective*. London: Oxford University Press

Gulliksen, H. 1987. *A Theory of Mental Tests*. Hillsdale, New Jersey: Lawrence Erlbaum.

Harris, D.P. 1969. *Testing English as a Second Language*. NY: McGraw-Hill.

Heaton, J.B. 1975. *Writing English Language Tests*, London: Longman.

Hymes, D.1972. On communicative competence. In Pride, J.B. and J. Holmes (eds.), *Sociolinguistics*. Harmondsworth: Penguin.

Lado, R. 1961. *Language Testing.* London: Longman.

Maley, A. 1986. A rose is a rose, or is it?: Can communicative competence be taught? In Brumfit, C. (ed.). 1986. *The Practice of Communicative Teaching*. ELT Documents 124. Oxford: Pergamon Press.

Morrow, K. 1979. Communicative language testing: revolution or evolution? In C.J. Brumfit and K. Johnson (eds.). The Communicative Approach to Language Teaching. Oxford: Oxford University Press.

Munby, J. 1978. *Communicative Syllabus Design*. Cambridge: Cambridge University Press.

Nitko, A.J. 1983. *Educational Tests and Measurement, An Introduction*. NY: Harcourt, Brace, Jovanovich.

Oller, J. W. (1991) Foreign Language Testing, Part 1: Its Breadth, *ADFL Bulletin*, 22: 33-38

Richards, J.C. 1990. Communicative needs in foreign language teaching. In Bolitho, R. and R. Rossner (eds.) 1990. *Currents of Change in English Language Teaching*. Oxford: Oxford University Press: 48-58.

Richards, J.C and T. Rodgers. 1988. *Approaches and Methods in Language Teaching*. Cambridge: Cambridge University Press.

Roberts, J.T. 1982. Recent developments in ELT. In Kinsella, V. (ed.). 1982. *Surveys 2*. Cambridge, Cambridge University Press.

Spolsky, B. 1975. Language testing: Art or science? Paper presentation. Stuttgart: fourth AILA International Congress.

Swan, M. 1990. A critical look at the communicative approach. In Bolitho, R. and R. Rossner (eds.). 1990. *Currents of Change in English Language Teaching*. Oxford: Oxford University Press: 73-98.

Valette, R.M. 1967. *Modern Language Testing: A Handbook*. NY: Harcourt, Brace and World.

van Ek, 1975. *Threshold Level*. Strasbourg: Council of Europe.

van Ek, J.A. and J.L.M. Trim. 1991. *Threshold Level 1990*. Strasbourg: Council of Europe.

Weir, 1990. *Communicative Language Testing*. New York: Prentice Hall

Wilkins, D.1976. *Notional Syllabuses*. Oxford: Oxford University Press.

Wright, T. 1987. *Roles of Teachers and Learners*. Oxford: Oxford University Press.

Xiaoju, L. 1990. In defence of the communicative approach. In Bolitho, R. and R. Rossner (eds.) *Currents of Change in English Language Teaching*. 1990. Oxford: Oxford University Press: 59-72.

Module 2

MODULE 2

THE TEST PRODUCTION PROCESS

All tests, whatever their purpose or level, must meet certain criteria, and it is appropriate, before looking in detail at how tests are produced, to look at these. The questions asked here underpin the whole process of test production.

CHECKLIST FOR THE EVALUATION OF TESTS

Test Validity

What is the purpose of the test?
Is the content of the test consistent with the stated goal for which the test is being administered - is it valid?

Test Difficulty

What are the characteristics of the examinees?
Is the test appropriate to the overall abilities of the examinees?
Has it been tried out on a sample of persons from the same general population as the target examinees?

Test Reliability

Are the test results reliable enough to make accurate decisions?
Can the degree of reliability be demonstrated, and how reliable is this test?

Test Applicability

How suitable is the test format and features to the context of use?
How familiar is the actual test format to the examinees?
Can the format and features of the test be fairly applied in the real testing situation?

Test Relevance

How relevant is the test to the proposed test population and/or to the test item domain?
How closely does the proposed test population/content resemble the developmental sample?

Test Replicability

How easy is it to produce equivalent or equated forms of the test?

Test Interpretability

How easy is it to score the test, report the test scores and interpret them?
Does it require a great deal of training?

Test Economy
What does it cost to procure, administer and score the test?

Test Availability
Is the test readily available?

Test Acceptability
Is the test societally and institutionally acceptable?
Is it acceptable in the eyes of teachers, parents, administrators?

Adapted from Grant Henning, A Guide to Language Testing, Newbury House, 1987.

In describing the test production process it is important to recognise that people who write items which appear in tests or on examination papers are involved in this process to a varying extent. This may range from responsibility for the production of a test at every stage to involvement only as a freelance writer who is asked to produce items, and possibly to take part in editing them.

It is therefore essential to consider not only the process of producing examination papers, but also the development of the examination itself. Each is a cyclical process, involving the feeding back of knowledge and experience gained, and the continuous re-assessment of the examination and each administration of it.

A model for the process of developing a new examination can be illustrated by figure 1. This diagram gives a blueprint for the stages to be gone through, starting from the initial perception that a new test is necessary, before the operational phase, in which item writers are most heavily involved, is reached. A more detailed example is given as appendix D.

As figure 1 shows, the test development model is cyclical. The need for a new test, once perceived, leads into a planning phase, during which data on the exact requirements of candidates may be collected by means of questionnaires and consultation with schools and colleges which are likely to use the test. A clear picture of who the potential candidates are likely to be should emerge; during the design stage, which follows, an attempt is made to produce the initial version of specifications of a test which will be suitable for those candidates. The appearance of the test and all aspects of its content are discussed, together with all considerations and constraints which affect this. Initial decisions are 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. Sample materials are written.

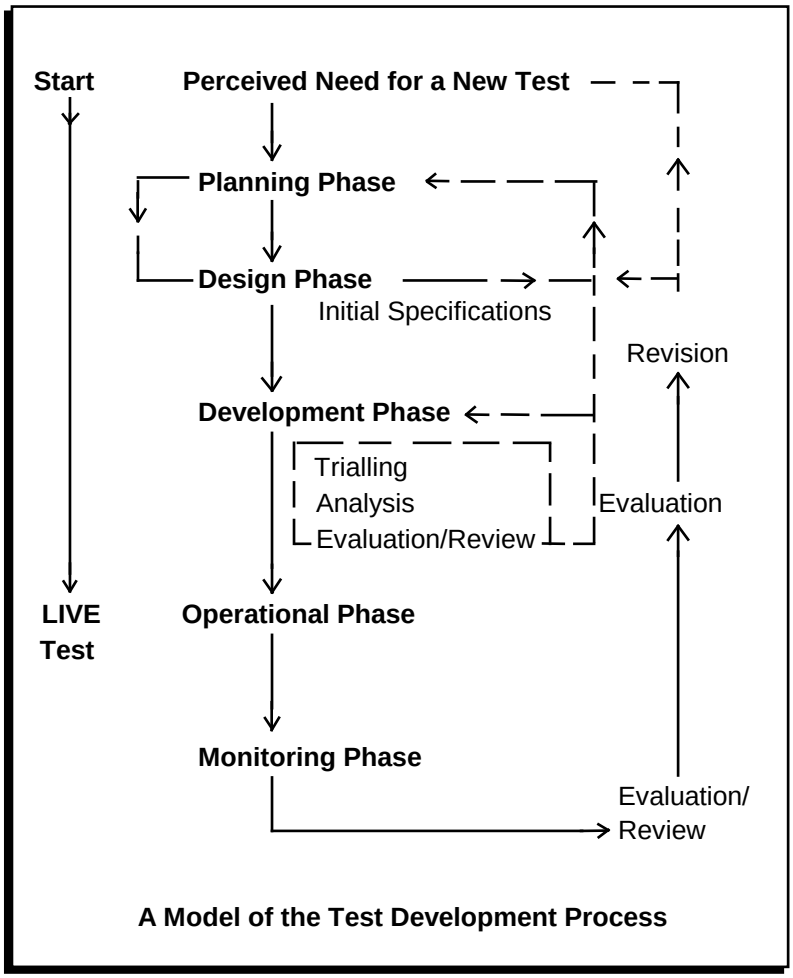


Figure 1. A test development model (Saville and Milanovic)

During the development phase these materials are trialled. This means that students who are at the appropriate sort of level to take the test and similar to projected candidates in terms of other factors such as age are given test materials under simulated examination conditions. Mark schemes and rating scales are devised. Information is gathered by means of questionnaires and reports from candidates and their teachers, video and audio recordings and observations, and score data is analysed. 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. At this stage it is still possible to make radical changes to specifications, the item types used, and in fact most aspects of the test.

Once these initial phases have been completed, the test specifications reach their final form, materials are written, and test papers constructed. A regular process of administering and marking the test is set up. This is the operational or 'live' phase, during which the test is available to candidates. This phase is shown in more detail in figure 3.

The results of test administrations are monitored, and feedback obtained from candidates and teachers at schools where the test is used. Such data is used in evaluating the test's performance and assessing 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. At some point revision will in any case become necessary, in order to reflect developments in testing theory and practice. This returns the test to the beginning of the cycle, as major revision of a test means going back to the planning phase.

Specifications

When the specifications for a new (or revised) examination are planned, the underlying aim is always to produce an examination which is **valid**, **reliable**, which has **impact** and which is **practical.** In other words, the test should offer an appropriate way of measuring what it claims to measure, the results produced should be as free as possible from errors of measurement, the effect it has on individuals and on classroom practice should be positive, and the demands it makes on the resources of the test developer should be compatible with the resources available. During planning these factors always need to be kept in mind, and a balance between them must be achieved.

The first stage of planning involves a situational analysis, which means looking at the need for a new examination within the context of all the various influences on it which will affect the form it finally takes, with the aim of identifying the principal considerations and constraints relevant to the project. These cover all aspects of what the examination must do in order to meet its purpose and the limitations placed on the examination by the circumstances in which it is to be used.

The considerations are of two types, 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.

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

- the number of staff and rooms available;
- how many candidates there are, and 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 include:

- the acceptability of the examination for all the people involved, candidates, their parents, teachers, owners of schools, etc.;
- the way the examination fits into the current system in terms of curriculum objectives and classroom practice;
- the level of difficulty required;
- external expectations of what an examination of this kind should be like, including comparisons which might be made with any similar products on the market;
- the availability of resources for the development, administration and reporting of results.

As figure 1 shows, when specifications have been drafted, a first attempt to design the test can be made, and sample materials produced. The materials are then trialled by administering them to candidates who are at an appropriate level, and the results analysed. In the light of trialling, item types or certain types of materials may be rejected, and the length of sections of the test or aspects of its administration changed. The specifications may undergo several revisions before they reach the form they are to take for the live test.

It may happen that the same person is responsible both for developing the specifications and then for writing materials for the live test. It must also be possible for people who have not previously been involved to get detailed information about the examination from the specifications, whether they need the information in order to decide whether to enter students for it, or in order to write items for it. An item writer who has not written for an examination 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.

The specifications which are finally produced should give detailed information on each paper or part of the test. They will cover the following areas:

- 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
- the focus of the item, e.g. showing detailed comprehension of a text
- what is being tested, e.g. use of grammatical rules
- text types used in input
- total and individual length in words of texts used
- text sources
- some indication of topic areas considered suitable for use
- 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 coordinated
- 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
- types of prompts used in tests of oral production
- details of grading procedures and reporting of results
- where and when the test can be taken
- availability of past papers
- estimated number of hours of study necessary as preparation for test

All this information helps to give item writers a very clear picture of the nature of the materials they are involved in producing.

Sometimes it is useful to be able to grasp the essential information about a test by means of a brief diagrammatic summary. An example of a presentation in the form of a grid of an examination which is made up of five 'papers', or components, is given below. Each component of the examination is summarized in terms of what the test focuses on, the input provided and the nature of the expected response.

Paper 1 - Reading Comprehension	***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 - Composition	***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. One question on each of the three background reading texts.	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 and written speeches.
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

The production process

The specifications give a definition of what must be produced for the examination; the actual process of production is likely to consist of five stages:

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

This process can be represented by figure 3 below. In looking at this diagram, it must be kept in mind that the model for tests which can be objectively marked varies somewhat from the model for the necessarily more subjective tests of speaking or 'free' writing, which cannot be pretested in the same way.

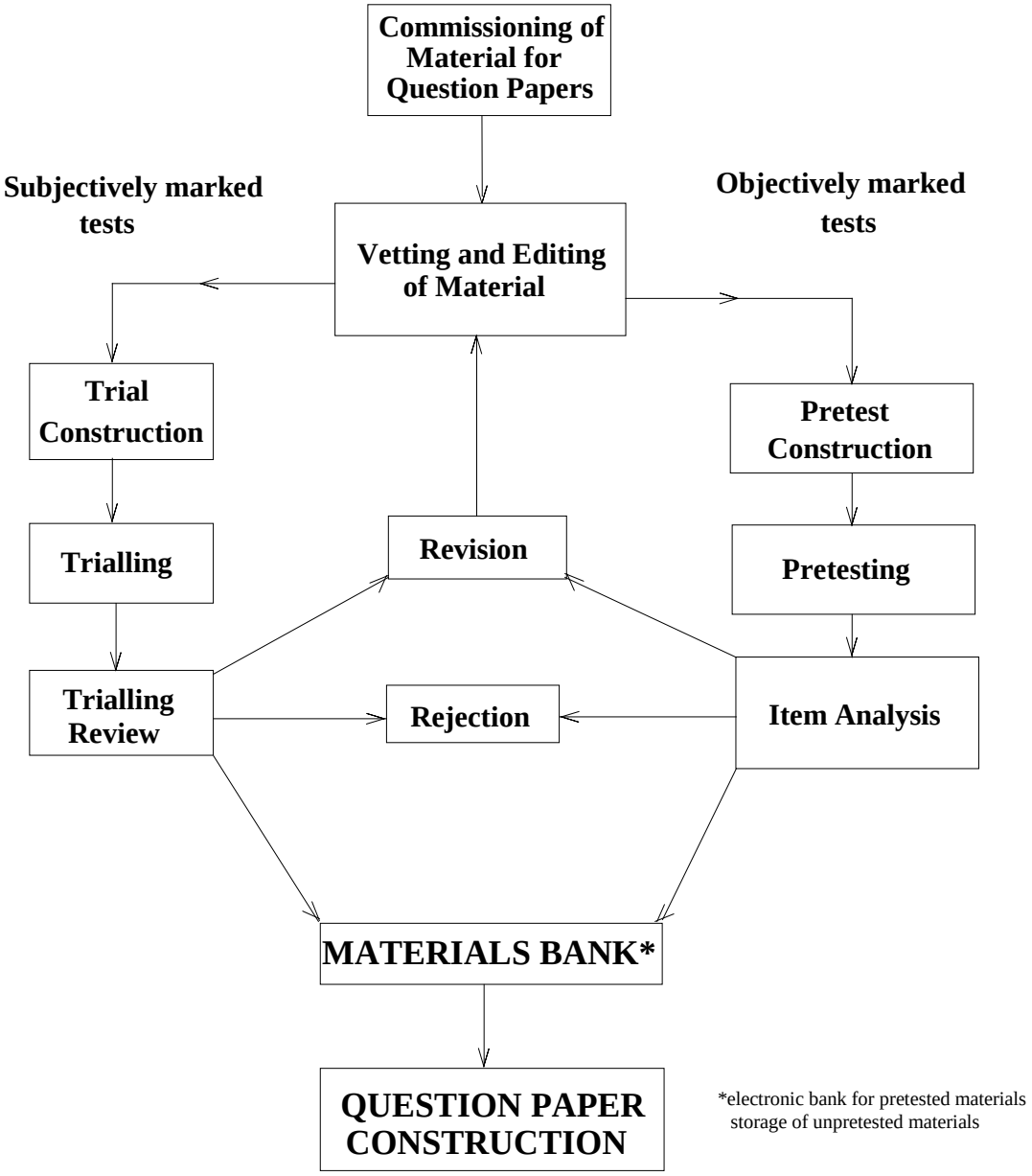


Figure 3. The operational phase of test production

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

- **scheduling**; this means drawing up realistic project plans, and then meeting deadlines set;
- **record keeping**; this is vital when any process involving revision and modification is concerned, and when materials may go through several versions.

a) Commissioning

As mentioned previously, the same person (who can be referred to as the examination developer) may be responsible for all stages of the production process, including writing test materials. However, at this stage it is common for the examination developer to commission a number of other people, either employees of the same organization, or outsiders connected with teaching or testing, who work as freelancers, to take part in selecting or writing texts, and in writing items.

It is sometimes the case that the same employee of the organization which produces any given examination is responsible for organizing the commissioning and editing stages in this process and for using the items produced in question paper construction, while others are involved in the pretesting, analysis and banking stages. The same person may be responsible for all parts of the examination, or in the case of an examination composed of separate tests of reading, writing, listening and speaking, for example, someone different may be in charge of each paper.

Commissioning may follow a regular pattern, and happen, for example, twice a year, or it may be done whenever the examination developer considers that new materials are needed.

Item writers may be asked to submit complete examination papers, or groups of particular items used in the exam.

The aim of the examination developer is to receive as high as possible a proportion of material which, after editing, will prove acceptable at the pretesting stage and will make its way into live papers. It is therefore part of their responsibility first to choose suitable people to commission as item writers and then to give them instructions which are as clear and helpful as possible. External item writers are often found among people who have some experience of the examination, either from preparing students to take it or from marking or oral examining. Whether the test developer is working alone or with colleagues or commissioning external writers, at this stage the following points need to be made clear:

- **Exact details of the materials required**. In the case of texts, it should be made clear whether items are to be written immediately, or only after acceptance of the text. The item writer should provide a key to all items. A recorded version of listening materials may be requested as well as a written script. 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 presentation of materials expected**. Handwritten copy may not be considered acceptable, but copy may be welcome on disk as well as on paper. 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 may be helpful if item writers put their name, the date and the name of the examination or examination paper on each sheet of paper. (All this may be covered in the internal specifications or other guidelines for item writers.)

- **Details of the deadline by which materials must be submitted.** It is useful for external writers to know how their job fits into the overall production schedule, so that the importance of keeping to deadlines is clear. When commissioning, it is usually possible also to give item writers some idea of when editing will take place, and either to let them know that they will be expected to take part in editing, or to ask them whether they wish to be involved.

- **Details of fees to be paid.** It should be made clear from the start to item writers what terms they are agreeing to work on. There may be a fee for accepted materials only, with no payment for any rejected materials, or a small fee may be paid on submission, 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 examination.

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

- specifications (internal or external: see below);
- sample materials or past papers;
- a handbook (or instructions or guidelines) for item writers for the specific examination or examination paper in question.
- a form on which to indicate acceptance of the commission;
- a form on which to indicate that the organization 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 organization which produces the examinations.

Specifications in their **external** form are available for public information, and, while they give a great deal of detail on the content of an examination, they would never include details of test production or of particular problems that involves. There may, however, also be an **internal** version of the same

specifications, which is normally a confidential document, and includes additional advice and guidance for the item writer.

The internal specifications should contain advice on the selection and presentation of materials which can prevent item writers from wasting time by making their own, possibly mistaken, assumptions about what is acceptable. Advice could take the following forms:

Advice on choice of texts

This 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.local papers)
- a general warning to avoid cultural bias
- a list of reasons why texts have been rejected in the past

Reasons for rejecting texts include:

- too great an assumption of cultural knowledge
- 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 idiosyncracies
- 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.

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 can best be illustrated with an example. The task is a modified cloze, designed to focus on words of a structural rather than lexical nature.

The following advice is given:

- 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 getting the next one wrong, too (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 'very') should be avoided, as should items which focus on very unusual or idiosyncratic structures.

The standard rubric to be used on the examination paper is given. This is followed by a checklist which can be used to check materials before submitting them:

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:
Have the required number of items been generated?
Are the items spread evenly through the text?
Is a good range of pronouns, conjunctions, etc. included?
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?

Some examples of rejected texts, together with a breakdown of why texts and individual items were considered unsatisfactory, are then given.

Having assimilated all the information and advice available, the item writer then has to produce the materials and meet the deadline. Writers who are anticipating being commissioned to produce materials for an examination which includes text-based items usually collect texts from the sources recommended as suitable, and choose the most promising ones to work on when the commission comes. For writing some types of items it is useful for the item writer to have a dictionary and thesaurus to hand.

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

Before submitting materials, item writers should check that they have kept a copy of everything, and that if texts such as newspaper or magazine articles are used, they have kept photocopies if the originals have been given to the examination developer.

b) Vetting and editing

When materials have been received from all the item writers commissioned, some preliminary decisions may be made on which materials should go forward for detailed editing, and which should be rejected immediately. This may be done by internal employees of the examining organization, or with the advice of external advisers

It is at this stage that texts which are clearly unsuitable for any of the reasons given above can be rejected. If texts without items have been commissioned, item writers may be asked at this stage to produce items on texts which have been accepted at vetting. It is always advisable for item writers who are asked to submit texts without items to have written the items, at least in a rough or preliminary form, in advance, so that if the texts are accepted, the items can be supplied quickly.

External item writers are not normally involved in vetting, but they are often asked to take part in editing the materials commissioned, and paid a fee which covers both time spent in editing and time spent studying materials in preparation for editing. The examination developer will decide:

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

Materials for editing should be sent out to those who are to attend the meeting in advance, so that everyone has time to work through them before the meeting. As well as reading texts in order to check their length, suitability of topic and style and level of language, it is also advisable to cover the key (if it is sent out) and work through the items as if taking the examination. This should make it possible to identify, for example, multiple choice 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 is likely to choose it, or items which are difficult or unclear even to a native speaker. **These materials should always be regarded as confidential.**

New item writers may be trained in editing by working in a group with more experienced people. 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.

In each editing group one person has to take responsibility for keeping a record of all decisions made about materials, clearly showing changes made at editing. It is unusual for materials to be accepted exactly as they were submitted, and common for accepted materials to be amended quite extensively in the course of an editing meeting. There is often a lot of discussion about materials, and, especially if they are text-based and at quite a high level, different opinions of their suitability can be defended. It is useful for the examination developer or somebody else with some degree of authority over the group to be able to make final decisions and decide when there has been enough discussion.

Opinions differ on whether item writers should edit their own materials. Some examination developers prefer to avoid this if possible, but if there are only a small number of people involved in editing, it may be necessary for everyone to take part in editing their own materials. From the point of view of getting immediate feedback and being able to avoid making the same mistakes in future, it is arguable that it is better to get other people's reactions at first hand, even if it is difficult to be present when materials you have spent a long time working on rejected.

At the end of the meeting, it is vital that there should be no future doubt about what changes were agreed on. A clear record of changes made to accepted materials must be kept. Some materials may appear to have some potential, but only if they are amended to an extent which could not be done in the course of the meeting. They may be given back to their original writers for further work (normally at no further fee) or may be given to a more experienced writer for revision and further editing. It should be made clear whether materials which have been rejected for some reason can then be taken back by their writers and possibly used for some other purpose such as classroom teaching, or whether they remain the property of the examining organization. The examination developer may wish to keep some examples of rejected materials for use in training sessions.

After the meeting, spare and used copies of the materials which have been edited should be destroyed. The amended copies of accepted materials are kept by the examination developer, who will usually have them re-typed in preparation for pretesting.

Item writers are entitled to expect some feedback from the examination developer on rejected items, especially if they have not been invited to take

part in editing, or if they have not edited their own materials. This helps them to avoid repeating the mistakes made this time when they submit materials in future.

At this point the direct involvement in the examination production process of many item writers usually ceases. When materials are typed up, names of item writers and details of text sources, etc. are no longer included.

Pretesting

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 used in a live test.

The more objective item types such as multiple choice and gap-filling are normally pretested. It is the individual items which are being tested, not the test as a whole, so a pretest paper does not necessarily resemble the live examination, for which the material was written, either in length or in composition. For example, pretests of items for an examination which consists of twenty discrete multiple choice items followed by three texts, each with five multiple choice items, may consist of forty discrete multiple choice items each and one text with five items.

Pretest papers are administered in the form of mock tests under simulated examination conditions to students of the language who appear to their teachers to be at the right sort of level to take the examination. They benefit by the exam practice and feedback on their performance which they get, while the examining organization, having had the pretest papers marked in the usual way, gets the data provided by statistical analysis of the items. In order for the data to be meaningful, it is normally thought that a pretest needs to be administered to at least 100 students.

The freer, more subjective types of tests, such as compositions, cannot be pretested in the same way as items for which there is one correct answer. However, a form of check can be made on these item types before they are used in live examinations. They can be **trialled**, again by being administered to volunteer students who are at about the correct level for the test. The answers they produce are marked by examiners who are used to marking the live papers, applying the usual criteria. It can then be seen 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 structure and vocabulary expected of candidates taking an examination at this level.

Item analysis

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 use of poor or faulty items in live administrations. However, it is important to realise that it is possible for a poor item to produce acceptable statistics, and to regard the results of this type of analysis as only one among the factors which determine 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. From this kind of analysis, information about the performance of individual items can be obtained. This is of the following kinds:

a) Item facility

It is necessary to know this in order 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. (In Figure 4, the attached MicroCAT printout, this statistic is given in the **'Prop. Correct'** column.) The appropriate level for the test is at the mid-point of the difficulty range, but an acceptable range of item facility might be set at .33 to .67 or .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 different facility level might be looked for in a test of proficiency to be given at the end of a course of study from that looked for in an aptitude test.

A test should include some items at each extreme of the range; in particular it is usual to provide some easy items at the beginning of a 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, although, assuming an item banking system is in place, they may be banked and considered for use in a test at a different level.

b) 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 used; the discrimination index and the point biserial correlation (the columns headed **Disc. Index** and **Point Biser**. on the MicroCAT printout, Figure 4) .

i. Discrimination index

Once a test has been administered to a number of candidates the candidates should be ranked (put 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.

Figure 14

MicroCAT (tm) Testing System
Copyright © 1982, 1984, 1986, 1988, 1993 by Assessment Systems Corporation

Item and Test Analysis Program – ITEMAN (tm) Version 3.50

Item analysis for data from file C:\ITEMAN\13002D93.D93 Time: 15.59

Item Statistics Alternative Statistics

Seq No.	Scale-Item Key	Prop. Correct	Disc. Index	Point Biser		Prop. Alt.	Endorsing Total	Endorsing Low	Point High	Biser.
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	

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 are counted, and:

nH = the number of candidates in the high ability group who answered the item correctly

and nL = the number of candidates in the low ability group who answered the item correctly

The discrimination index, di, can then be defined as;

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

.

di 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 0.3 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.

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 0.3 are considered acceptable.

When a negative point biserial correlation appears it means that strong candidates failed to choose the correct answer for the item. This may show that an option other than the intended one can legitimately be seen as the correct answer. This 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 and pretest it again.

Distractor tallies

This kind of 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).

On the MicroCAT printout, the proportion of candidates choosing each distractor is given in the **'Prop. Total'** column:

To give a couple of examples:

a 'Prop. Total' column which shows	A	.15
	B	.10
	C	.63
	D	.12

where C is the key, shows an item where key and distractors are all performing satisfactorily;

a 'Prop. Total' column which shows:	A	.95
	B	.04
	C	.01
	D	.00

where A is the key, shows an item which almost every candidate could answer correctly, and where one of the distractors was so weak that nobody chose it.

Columns headed **'Seq. No.'** and **'Scale-item'** also appear on the printout. These refer to the item's **sequence number within the data set** and **the number of the scale that the item was assigned to and the item's position within that scale.**

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	Number of items included in the analysis.
N of Examinees	Number of candidates included in the analysis.
Mean	Dichotomously scored items - the average number of items that were answered correctly. Multipoint 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	Gives idea of the shape of the distribution.
Kurtosis	The peakedness of the distribution.
Minimum	Lowest candidate score.
Maximum	Highest candidate score.
Median	Middle candidate score.
Alpha	Alpha reliability coefficient for each scale ranging from 0.0 to 1.0. Is an index for the homogeneity of a scale. Ideally the value should be as close as possible to 1.
SEM	Standard Error of Measurement. 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 score is between 61.4 and 72.6.

Mean P The average proportion of correct answers (dichotomous items only).

Mean Item-Tot. Average point biserial across all items in the scale (dichotomous items only).

Mean Biserial 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%).

Figure 15

MicroCAT (™) Testing System

Item and Test Analysis Program - - ITEMAN (™) Version 3.50

Item analysis for data file C:\ITEMAN\13001D93.DAT

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	5	6	7
N of items	5	10	10	10	6	5	15
N of Examinees	270	270	270	270	270	270	270
Mean	3.230	6.633	8.422	8.163	3.778	1.959	12.807
Variance	0.725	3.321	1.755	2.588	1.751	1.321	3.259
Std. Dev.	0.851	1.822	1.325	1.609	1.323	1.149	1.805
Skew	0.047	-0.348	-0.361	-0.709	-0.288	0.402	-0.622
Kurtosis	-0.491	-0.202	3.043	-0.148	-0.325	-0.153	-0.292
Minimum	1.000	1.000	2.000	3.000	0.000	0.000	7.000
Maximum	5.000	10.000	10.000	10.000	6.000	5.000	15.000
Median	3.000	7.000	9.000	8.000	4.000	2.000	13.000
Alpha	0.091	0.431	0.318	0.499	0.371	0.407	0.570
SEM	0.812	1.375	1.094	1.138	1.050	0.885	1.183
Mean P	0.646	0.663	0.842	0.816	0.630	0.392	0.854
Mean Item – Tot.	0.428	0.406	0.378	0.415	0.493	0.541	0.342
Mean Biserial	0.676	0.547	0.602	0.621	0.662	0.753	0.590
Max Score (Low)	3	6	8	7	3	1	12
N(Low Group	168	116	115	89	115	103	105
Min Score (High)	4	8	9	9	5	3	14
N (High Group)	102	85	155	132	89	84	109

N (Low) The number of candidates in the lowest scoring group. This will be approximately 27% of the total sample.

Min Score (High) The minimum score a candidate could attain and be in the high ability group.

N High The number of candidates in the highest scoring group. This will be approximately 27% of the total number of candidates.

Item calibration

The type of classical item analysis described above is useful, but what it reveals about item performance can only be interpreted in relation to the candidates who took that particular pretest. It may, of course, be possible to make generalizations about the level of the material, especially if a good deal is already known about the pretest population. It is, however, difficult to make exact comparisons between items which have been pretested under different circumstances, and this is why Rasch analysis has been developed, as an extra statistical approach to classical item analysis. By including common items in each pretest and then determining the difficulty of the pretest items in relation to these common items, it is possible to create a difficulty scale. Essentially, item calibration involves putting items from different pretests with different candidates onto this one difficulty scale.

In order to make it possible to calibrate items, the difficulty scale has first to be established for a live examination administration, and then **anchor items** from this exam are included in pretests of new items. The anchor items reveal the relative difficulty of new items, putting them on the same scale. When a live test is compiled from calibrated items, scores in this test can be interpreted in terms of grades even before the test is administered.

The Rasch analysis of data makes a system of **computerized item banking** possible. This may be the next stage in the test production process. If Rasch analysis is not applied to items and a computerized banking system is not available, pretested items which are not rejected or in need of revision are simply held by the examinations officer, ready for use in constructing live examination papers.

Establishing a difficulty scale

Institutions which regularly administer tests will inevitably want to make sure that the grade boundaries they set are maintained from one administration to the next. Traditionally this is done through the grading process which first identifies whether or not the population has changed in any way since the last administration and then locates appropriate grade boundaries, taking account of the difficulty of the exam and the overall ability of the examination population.

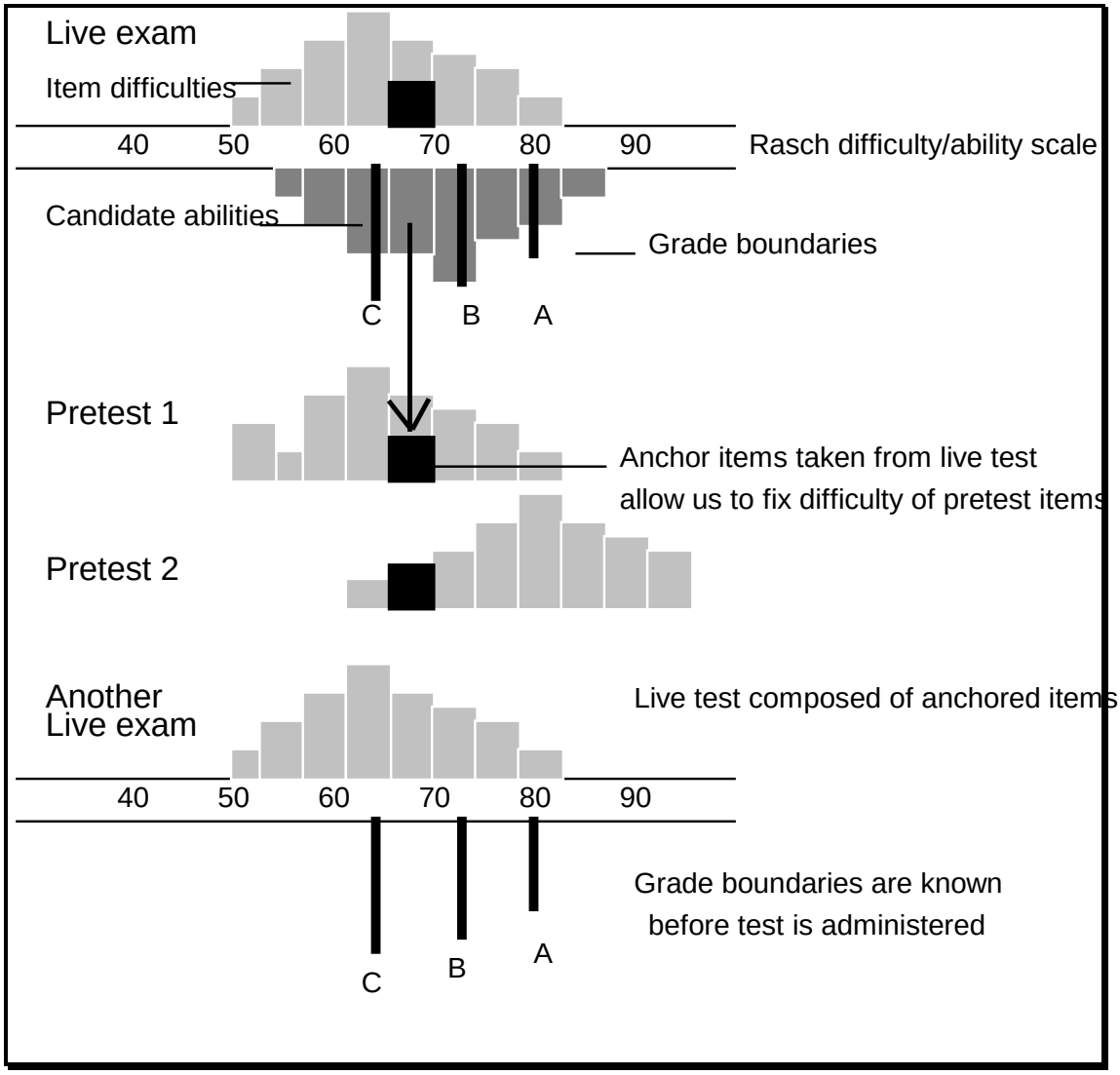


Figure 6. Item calibration

It is often the case, however, that the difficulty of tests needs to be known in advance to ensure that successive versions of a test are equivalent and are therefore fair to all candidates who sit the different versions. Even allowing for the grading process, it is still desirable for tests to have a similar statistical profile, as institutions need to be confident that candidates resitting their exams are likely to improve their grades, assuming that they have followed a course of instruction and are not affected by any adverse conditions such as illness or nerves. For these reasons, it is important to identify the overall ability of an examination population in relation to a representative set of items. This is done by analysing a set of data for a 'live' test involving about 300 candidates using a Rasch analysis package. This will estimate the difficulty of all of the test items in relation to each other. Clearly the most difficult will be the item which fewest people answer correctly, and so on. The difference between this approach and classical item analysis is that Rasch analysis ranks the items in terms of the *probability* of candidates getting an item

correct. In so doing, it reports item difficulties on a probability scale which uses a measurement unit called the *logit*. The word logit is derived from the logarithmic process that is used to estimate item difficulty. The difficulty scale starts at zero and then reports item difficulties which are either positive (more difficult) or negative (easier). So for example, if item 1 has a difficulty of 4.2 and item 2 has a difficulty of -0.2, one can see straight away that the second item is easier than the first.

Item difficulty can only be determined within the context of candidate ability. Accordingly, Rasch analysis places the item difficulties on a scale where they are relative to each other and then locates the test candidates on the same scale according to their ability. Once again, the most able candidates will have a higher reported ability so, for example, a candidate with an ability of 2.3 is, on this test, more able than a candidate with a reported ability of -1.5. This information is presented in the following way:

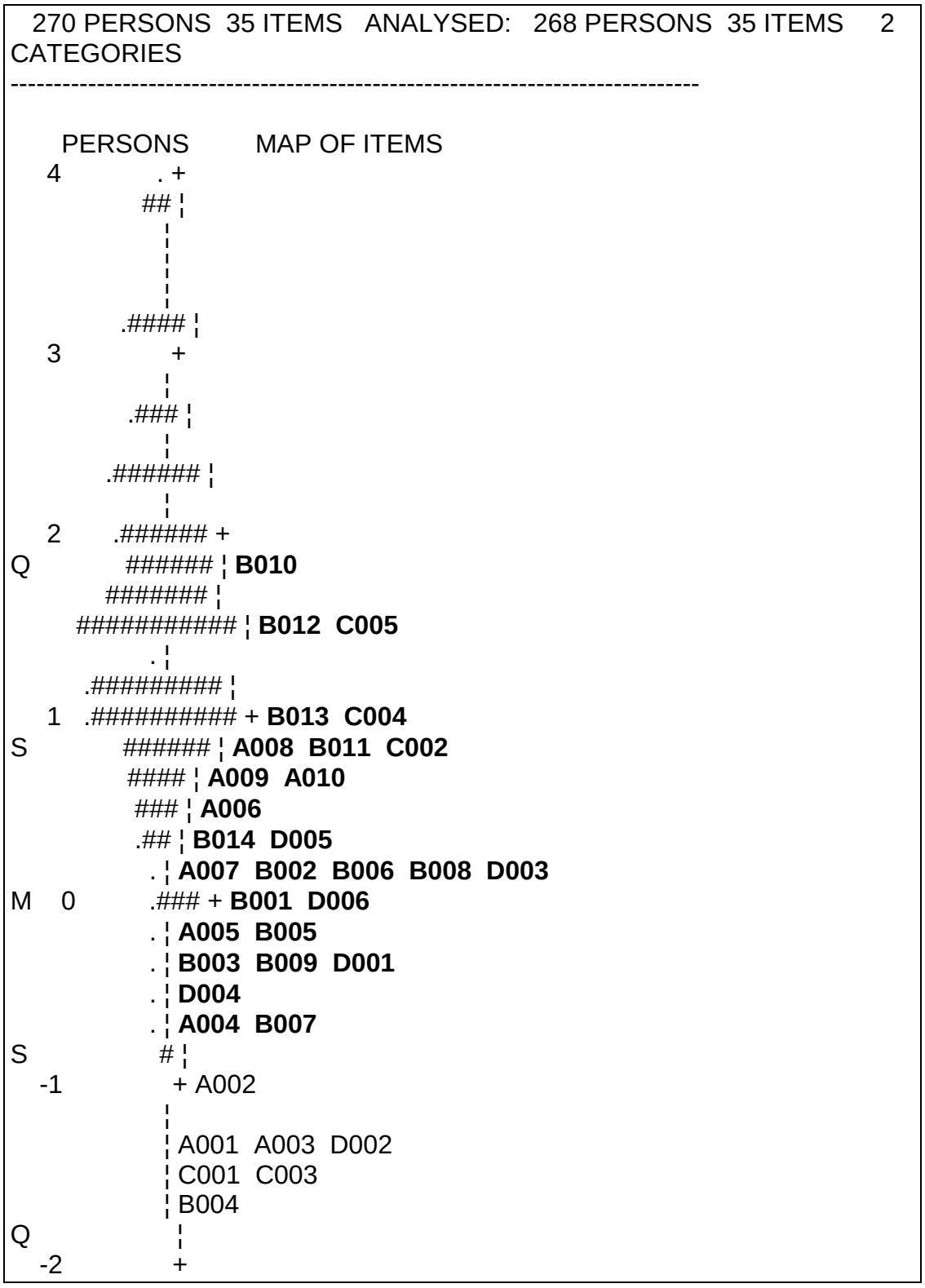


Figure 7. A Rasch scale

This table shows the items on the right in relation to the ability of the candidates on the left. Item 10 is the most difficult with item 4 the easiest. The letters which precede item numbers simply refer to the group of items, as

on some tests you may wish to code items according to groups or sections. For this particular test, some of the items were too easy for the candidates who took them. The items highlighted are at an appropriate level with any items below -1 logits being too easy. This is because the mean ability of this group of candidates is 1.45 whereas the mean ability of the items is, as always, zero. This difference of two logits means that for the average candidate there is an 80% probability of success on these items because as the difference in ability measure and item difficulty increases, then the likelihood of success on an item increases. The following table demonstrates how ability measure can be interpreted.

Logit difference between ability measure and item calibration	Probability of success on a dichotomous item	Logit difference between ability measure and item calibration	Probability of success on a dichotomous item
5.0	99%	-5.0	1%
4.6	99%	-4.6	1%
4.0	98%	-4.0	2%
3.0	95%	-3.0	5%
2.2	90%	-2.2	10%
2.0	88%	-2.0	12%
1.4	80%	-1.4	20%
1.1	75%	-1.1	25%
1.0	73%	-1.0	27%
0.8	70%	-0.8	30%
0.5	62%	-0.5	38%
0.4	60%	-0.4	40%
0.2	55%	-0.2	45%
0.1	52%	-0.1	48%
0	50%	0	50%

Figure 8. Logit-to-Probability Conversion Table

Ideally, the distribution of item difficulties will mirror the candidate abilities if the test is to be considered appropriate in terms of the degree of difficulty.

When the analysis has been completed and the item difficulties known in relation to the examination population, then this test can be used as a fixed point onto which new tests can be anchored. This can be done by using certain items from the test as anchor items.

Anchoring

Anchor items are generally added to different pretests as a way of ensuring that all new test material can be placed onto a common scale. It is assumed that the original test, the fixed point onto which future tests will anchor, is secure and will not be used for practice material and that candidates will not have been allowed to take copies out of the exam room.

The procedure for anchoring is as follows:

- Identify a set of items that are appropriate in terms of difficulty and fit (the fit statistic being determined as an indication of discrimination. Ideally the mean fit statistic for a set of items will be between +2 and -2 logits);
- Include these items in all pretest versions;
- Make up an anchor file with a list of the difficulty values in logits for the items you have chosen for your anchor section. Number these to correspond with their position in the pretest and save as an ASCII file;
- Indicate in the Rasch control file the name of the anchor file (IAFILE = name of file and directory etc. *a:/anchor.anc*)
- Run the analysis which will produce item difficulties that no longer have a mean of zero because they are now measured in relation to the ability of the original examination population.

Some institutions may not have the resources to include pretesting in their test development cycle. In these situations it is still possible to adopt an approach using Rasch analysis by including anchor items from the previous 'live' administration.

Score interpretation

Score interpretation using Rasch analysis is generally through the use of the Raw Score to Ability Transformation tables (Table 22 of the Bigsteps output file) that are generated by the analysis. The value of this table is that it solves the problem of deciding how to interpret raw scores on different versions of a test. Each raw score mark requires a certain level of ability in order to achieve that mark; the higher the mark, the higher the ability required to achieve it. Once tests are on the same scale, it is simply a question of identifying the ability measure required for candidates at a particular point, for example 60/100, then looking at the same table for the alternative version of the test and finding the ability measure previously identified. The raw score achieved by candidates of this ability will be the raw score equivalent of 60/100 for the first test. A Raw Score to Ability Transformation table is shown below for a 48 item test. It can be seen that if a pass mark of 30 was decided upon for this test it would mean that subsequent versions would need to set their pass marks at whatever raw score corresponded to an ability of .71. This is

assuming that anchor items are used to link all test items onto one difficulty scale using the procedure outlined above.

120 PERSONS 48 ITEMS ANALYSED: 120 PERSONS 48 ITEMS 2 CATEGORIES

TABLE OF MEASURES ON COMPLETE TEST

SCORE	MEASURE	S.E.	SCORE	MEASURE	S.E.	SCORE	MEASURE	S.E.
0	-5.45E	1.43	17	-.82	.35	34	1.21	.36
1	-4.72	1.03	18	-.70	.35	35	1.35	.37
2	-3.95	.75	19	-.58	.35	36	1.49	.38
3	-3.48	.63	20	-.46	.34	37	1.64	.39
4	-3.12	.56	21	-.34	.34	38	1.80	.40
5	-2.83	.52	22	-.23	.34	39	1.97	.42
6	-2.58	.48	23	-.11	.34	40	2.15	.43
7	-2.36	.45	24	.01	.34	41	2.35	.45
8	-2.16	.43	25	.12	.34	42	2.57	.48
9	-1.97	.42	26	.24	.34	43	2.82	.52
10	-1.80	.40	27	.35	.34	44	3.11	.56
11	-1.65	.39	28	.47	.34	45	3.47	.64
12	-1.49	.38	29	.59	.34	46	3.95	.76
13	-1.35	.37	30	.71	.35	47	4.72	1.03
14	-1.21	.37	31	.83	.35	48	5.45E	1.43
15	-1.08	.36	32	.95	.35			
16	-.95	.36	33	1.08	.36			

Figure 9. A Raw Score to Ability Transformation table

If items have been pretested with anchor items, then it is possible to separate them from their original pretest and store them in an item bank. With the creation of a well-stocked item bank, the potential for test construction increases enormously.

Item-banking

An item bank is very like a database. It is a way of storing test items in an organized manner, making it possible to choose items which match the precise specifications of a particular examination both easily and quickly.

One great advantage over the traditional way of producing and storing test materials is that, whereas in the past materials were both commissioned and stored solely for use in one specific examination, an item bank can hold items which have been commissioned for many different tests and different levels. It is therefore possible to devise a new test, possibly at a level midway between two existing ones, or with a range wider than that of an existing test, and go to the item bank for the range of items already in existence at an appropriate level of difficulty.

However, the potential of item banking goes beyond efficiency considerations alone. Item banking also makes possible a *qualitative* improvement on traditional approaches to test construction because the items in a bank are calibrated. This means that the difficulties of all the items are expressed on a single measurement scale. A candidate's performance on a test generated from an item bank can be reported as a location on this single scale, which simultaneously defines item difficulty and person ability.
By enabling us to locate learners accurately on a single proficiency scale, as we do subjectively anyway, using terms like 'beginner' and 'lower-intermediate', item banking goes some way towards overcoming the problem of interpreting performance on different tests. In other words, it offers a practical solution to the problem of test equating.

An item bank also has the potential to throw light on the nature of an ability being tested. The items, ordered as they are by difficulty, constitute a very detailed description (in terms of tasks that learners can typically perform at different levels) of the trait being tested. Thus, item banking has applications in research, and for construct validation studies.

Item banking and test construction

The three requirements for an item banking approach to test construction are:

- a system for collecting information on the performance of items, so that they can be calibrated;
- statistical methods for estimating item difficulties, anchoring them to a single scale;
- a system for storing and retrieving information about items.

The first of these requirements is met by the process of pretesting and the second by Rasch analysis of the data this provides. A storage and retrieval system necessitates the development of item banking software.

An example of item banking software

One item banking system (referred to as IBS) can be taken as a typical example of what item banking software has to offer. It has the following main features:

- it is developed as a modern Windows application;
- it deals flexibly with a range of item types: any kind of discrete item, or items grouped by association with a text;
- the user can define *attributes* to describe items, developing a descriptive system of whatever complexity is required;
- a fully-developed query system allows sophisticated item searching;
- it produces a range of useful reports, including an Ability Report which equates scores in a test with Rasch ability estimates;
- it keeps records of history of item use.

The item types in a banking system

The IBS can distinguish three generic item types. Figure 6 shows schematically these three item types:

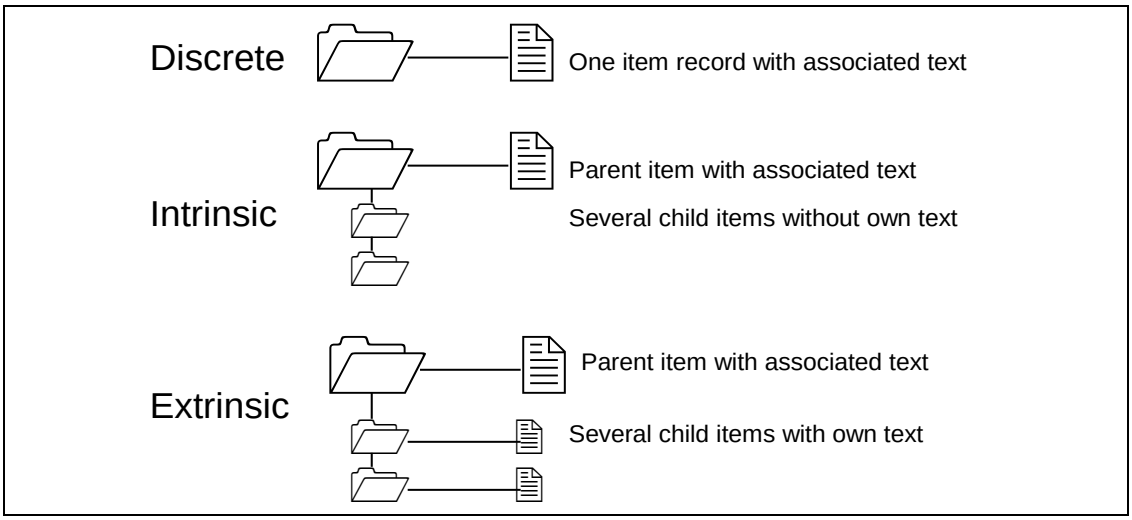


Figure 10. Three formal item types handle all possible situations

An example of an intrinsic item would be a cloze passage. Each gap in the text is a sub-item (or child) which has its own record in the bank, including its own difficulty value.

An example of an extrinsic item is a reading passage followed by comprehension questions. In this case the questions are associated with the text but potentially detachable.

The attribute system

Attributes are the characteristics used to describe materials in tests. They can refer to items, passages, graphics or any other characteristic of a piece of material that needs to be documented. Attributes can be numbers or words, and validation can be specified. For example, an attribute DISCRIMINATION could be specified as a decimal number with a value between -1 and +1. If this type of validation is specified, then only in-range values can be entered against items. Word attributes can be validated against a list, or left open-ended. For example, an attribute such as TOPIC could be restricted to a list of values such as Health, Leisure, Science, Politics, Other.

The query system

At the core of any item banking system is the query system. This allows for the selection and display of materials for the purposes of test construction or even research. It is important that the query system is as comprehensive as possible. Queries are built up using statements such as:

SELECT 10 ITEMS WHERE DIFFICULTY BETWEEN 50 AND 60 AND TOPIC EQ HEALTH

There is virtually no limit on the complexity of queries, and they can be saved and recalled. Successive queries can be used to add matching items to a list. Where more items are available than are asked for, the required number is selected at random. The user can then view the list of matches and the list of items actually used, and switch items between these lists.

Reports

When a test has been constructed, a number of reports can be produced. Two of the most useful are the **ability report** and the **distribution report**.

Grade	Score	Ability	Std Error
	29	97.82	9.32
A	28	91.27	6.76
B	27	87.25	5.65
	26	84.25	5.02
	25	81.81	4.60

C	24	79.72	4.30
	23	77.86	4.08
	22	76.18	3.91
	21	74.62	3.77
	20	73.15	3.67
D	19	71.75	3.59
	18	70.40	3.53
	17	69.09	3.49
	16	67.81	3.47
	15	66.54	3.45

Figure. 11. Part of the ability report for a 30-item test.

The ability scale is derived from the pretest statistics stored in the bank, and the standard error column represents the amount of error associated with different points on that scale. It is possible to define grade boundaries on the basis of ability levels, and, having constructed a test, the user can see at once how many points out of 30 must be scored to demonstrate a given ability and thus gain a particular grade. In other words, this test has been graded before being administered. One point worth emphasizing here is that grading prior to live examination use presupposes accuracy at the pretesting stage.

The **distribution report** gives a useful summary of the difficulty of the items used in a test. Figure 12 shows a part of this report for the same example test.

```
   TEST: FCE 1 (GRADES)


A  | 96 |
-----------------------------------------------------------
B  | 90 |
-----------------------------------------------------------
C  | 80 |
   | 79 |      i260
```

```
  | 78 |     i259      i258
  | 77 |     i253
  | 75 |     i250
  | 74 |     i242
  | 73 |     i237      i240
-------------------------------------------------------
D  | 72 |     i236
  | 71 |     i227      i233
  | 69 |     i220
  | 68 |     i218      i216
  | 67 |     i209      i210
  | 66 |     i203
```

Figure 12. Part of a distribution report

The distribution report lists items by difficulty against the same test scale as the ability report. The report shows how the items in a test are distributed.

Test construction without an item bank

It can be seen above how a system of computerized item banking can make test construction easier for examination developers.

If no such system is available, the examination developer will probably arrange a test construction meeting with one or two other people who know the examination well, possibly including an external adviser for the examination or examination paper.

The examination or examination paper is put together according to the specifications, using the edited and pretested materials resulting from past commissions. When this is done, certain considerations, have to be kept in mind, such as the range of text topics used throughout this examination, and also the range used in recent years, so that there is some variety within the scope of what is acceptable. Where discrete items are concerned, for example, a section of a test consisting of twenty multiple choice items which test various grammatical points, an attempt should be made to pick items which test a range of different points and also have some range of difficulty. If the examination is known to be taken by a wide population of candidates from many countries and in varying age groups, the overall choice of text topics

and items should not appear to favour the interests of any particular group, although obviously some candidates will always find certain texts more interesting than others.

Finally, the selected texts and items go through the process of design, layout and printing in the style established for the examination, and are held in secure conditions until their administration.

EXERCISES

1. The specifications are being planned for a high level (ALTE Level Four or Five, C1 or C2) test of speaking, to be used as part of a university entrance qualification for non-native speakers.

Look at the professional considerations listed on the next page.

What effect would you expect each of these considerations to have on the decisions you would make about the test?

2. A test of general language proficiency at an intermediate level (ALTE Level 3, B2) is being planned, which is aimed mainly at people in the 15 to 20 age-group who may be of any nationality.

Which topic areas would you consider suitable for use in reading and listening texts and exercises?

Which sources of texts could be suggested to item writers?

3. The texts given in Appendix A were all rejected at vetting when they were submitted for use in the reading subtest of an examination in general proficiency aimed mainly at young adults (who are probably working rather than studying). They might be of any nationality, and could be taking the examination in their own country or abroad.

Texts used in this examination are often lightly edited and simplified, but should not be heavily re-written. They are usually approximately 250-300 words long, and item writers are asked to produce five multiple choice items to accompany each text.

What reasons would you give for the decision to reject each of these texts?

4. The section of a MicroCAT printout given in Appendix B shows how some multiple choice items performed in a certain pretest. Say what you can about each one by ansering the following questions:

a) how easy does it appear to have been for the candidates?

b) how well did it discriminate?

c) did the key and distractors form a good set of options?

Suggested answers to exercises:

1. The first professional consideration concerns the types of real-life situations in which the candidates will need to use the language. In this case, they will need to be able to discuss academic matters with university staff and fellow students, which will involve expressing and justifying opinions, agreeing and disagreeing with other people with an acceptable degree of politeness. They will also have to deal with situations concerned with day-to-day survival in a foreign country (finding accommodation, shopping, etc.) and casual socializing with other students.

The second consideration concerns the level of performance necessary for the situations mentioned above. The ability to meet the demands of studying is most important here. If non-native speakers have to be able to participate in discussions with native speakers on academic topics, the level of performance will have to be high. A lower level might be sufficient for the non-academic needs mentioned.

The third consideration concerns the real-life language events which need to be re-created in the testing context. In this examination it will be most relevant to re-create situations involving argument, discussion, expressing opinions, agreeing and disagreeing, asking for and providing information. The test may take the form of paired interviews, so that there will be opportunities to devise tasks which involve candidates in discussion and reporting of each other's opinions.

The fourth consideration concerns the information to be given to users of the test. Whether or not a university using this test is explicit in letting candidates know exactly why they have failed to reach an aceptable level will have some connection with whether or not it is thought desirable for people to be able to prepare specially for the test.

2. Topic areas might include the following:

> family, friends and home life, hobbies and leisure interests, sport and games, studying and education, careers and future ambitions, entertainments and cultural interests (such as music, theatre, cinema, reading), holidays, travel, festivals and celebrations, clothes and fashion, shopping, food and cooking, buildings and places, some environmental and social issues e.g. those concerning tourism, city life, the countryside.

Sources of texts might be: magazines, brochures, newspapers, advertising literature, works of fiction and books on factual topics, radio programmes.

3. The following should be mentioned among the reasons for rejecting these texts:

Text 1- The topic is unsuitable, being of real interest only to northern Europeans. It is also likely to upset any candidate who has been affected (or had a relative affected) by cancer.

Text 2 - Texts made up of many small separate paragraphs like this can be used, but a lot of this text is taken up with addresses, and the information given may not provide enough points for testing above a very elementary level. The style is jokey and very journalistic, with a lot of playing with words in a way which may not be suitable for use with, or comprehensible to foreign learners - 'kiss-me-quick Newquay', 'an unbeatable natural high', etc. Light-hearted journalistic texts often prove unsuitable for these reasons, although they look superficially attractive.

Text 3 - The topic is not suitable for people from every culture; many are not familiar with the idea of animals as lovable pets, and regard dogs in particular as dirty pests. Other cultural problems are the idea of a society like the RSPCA, and giving an animal as a Christmas present.

The article also seems to be aimed at children and parents rather than at young people who are very likely to be between these categories in age.

Vocabulary like 'romping', 'cute, pudgy puppy', 'floppy ears, 'doggy best friend' would have to be modified.

Text 4 - Eating out may seem a suitable topic for this age-group, but this article is very culture specific. The idea that it is interesting to eat out in restaurants featuring food from various countries is not universal, and, where it is common, may be rather class-specific. There are many assumptions of shared cultural knowledge among the readers of a publication like 'Time Out', typified by a comment such as 'committed but not worthy', which is made about a vegetarian restaurant. Mention of alcohol is best avoided, if candidates from cultures where it is forbidden may take the test.

There are many foreign words such as 'sushi', 'sashimi', 'calzone', which might be seen as giving an advantage to some candidates and a disadvantage to others. Colloquialisms like 'pit-stop', 'booze' and 'shoestring' are a problem arising from journalistic style.

Text 5 - Sport seems a safe topic to use with this age-group, but this article is specialised to the point of being incomprehensible. It is also presented in an extremely colloquial style, which replicates speech rather than any common written style, and the humorous tone would probably only add to problems of comprehension.

Text 6 - Topics such as wildlife, investigations of the past and other areas of popular science are often suitable for use in examination texts. This topic may be considered to be non-controversial and of general interest, but will it yield enough multiple choice items? It does not really say anything beyond telling the reader that dinosaur footprints have been found. Despite some quite

difficult vocabulary such as 'dubbed', 'fared' and 'clams', the content is rather thin.

4. Comments on MicroCAT examples:

Item 1. This was an easy item (proportion correct = .92). It did not discriminate particularly well (point biserial correlation = .25, just into the acceptable range). The key and distractors did not form a very good set, as most people identified the key correctly, but distractor C attracted no candidates and distractor B only .01 of candidates.

Item 2. This was a relatively difficult item (proportion correct = .47). It discriminated well (point biserial correlation = .47). Between them, distractors B and C attracted more candidates than did the key, A.

Item 3. This was quite an easy item (proportion correct = .77), which did not discriminate very well (point biserial correlation = .29). One of the distractors (D) did not attract any candidates, which suggests that it was a weak, implausible distractor.

Item 4. This was quite a difficult item (proportion correct = .43), which did not discriminate well (point biserial correlation = .12, unacceptably low). Almost as many candidates (.40) chose distractor C as chose the correct option (.43), while B was a relatively weak distractor (.01).

Item 5. This was a difficult item (proportion correct = .38), which discriminated quite well (point biserial correlation = .33). More candidates chose distractor A (.41) than chose the key (.38), with some also choosing C and D. However, it seems to have been the weaker candidates who were misled, so the item appears to work well.

Item 6. This was quite an easy item (proportion correct = .83), which discriminated quite well (point biserial correlation = .37). Each distractor attracted some candidates.

APPENDIX A

REJECTED TEXTS

1. The Dangerous Sun

We have only recently become more conscious of the damaging effects of the sun. Experts believe that changes in the environment have made exposure to the sun more dangerous. Strong links have been found between thinning of the ozone layer and skin cancer. This thinning will mean that the screening out of the harmful UV radiation will be less successful. This is most alarming, since for every 1 percent increase in UV the incidence of skin cancer goes up by 2 percent.

The rate of skin cancer is increasing rapidly. Sunbathing, for example, is now seen as more hazardous than previously recognised. Health experts believe that increasing exposure to the sun without protection is the major factor in the increase. It has become increasingly urgent to educate the public about the risks of too much exposure to the sun.

Most skin cancers are curable. Those which are not are mostly in one of the three categories of skin cancers - malignant melanoma. Everyone should learn the early signs of skin cancer and if they are discovered a doctor should be consulted. Even with the most serious forms of cancer if diagnosed and treated early there is a very good chance of achieving a cure.

Children's skin is especially delicate and experts believe that long, unprotected exposure to the sun in childhood dramatically increases the risk of skin cancer. Since fifty percent of our lifetime exposure to UV radiation occurs in the first 20 years of life, high factor sun screens become extremely important.

The Australians are very aware of sun-related problems. They have the highest rate of skin cancer in the world and are very advanced in promoting sun awareness. In the UK, the medical profession and manufacturers of beauty and sun care products are helping in a campaign to educate the public about sensible sun protection.

2 A to Z of Activities

Hang Gliding
Hang-gliding courses provide an unbeatable natural high but young people must be over the age of 15 to take part.
Contact: The British Hang Gliding and Paragliding Association (BHPA) c/o Jennifer Burdett, Old School Room, Loughborough Road, Leicester LE4 5PJ. Tel: 0533 611322 or Fax: 0533 611323.

Initiative

Seize it!

Jousting
At a medieval castle near you - being a spectator rather than a participant might be advisable here!
Contact: the British Jousting Centre, c/o Max Diamond, Tapely Park, Instow, North Devon EX39 4NT. Tel: 0271 861200.

Karting
Go, go, go, Karting. Try life in the fast lane and remember this is how Mansell, Prost, Senna et al started their careers.
Contact: The RAC Motor Sports Association Ltd., Motor Sports House, Riverside Park, Colbrook, Slough SL3 OHG. Tel: 0753 681736.

Lifesaving
The Royal Life-Saving Society UK is the country's principal authority on lifeguard training.
Contact: RLSS UK, Mountbatten House, Studley, Warwickshire B80 7NN. Tel: 052.785.3945 or Fax: 052.785.4453.

Mountaineering
From Ben Nevis to Snowdon the challenge is steep.
Contact: UK Mountain Training Board, Capel Curig, Gywynedd, Wales. Tel: 06904.272.

National Trust
With over 600,000 acres of land and 535 miles of coastline, Britain's National Trust is the largest conservation charity in the world. The Education Group Membership provides resources and teaching as well as free access to many Trust properties.
Contact (for a free education pack): The National Trust, 36 Queen Anne's Gate, London SW11 9AS. Tel: 071.222.9251 or Fax: 071.222.5097.

Orienteering
Take a compass bearing and step this way for details of local clubs, activities and a list of publications.
Contact: British Orienteering Federation, Riversdale, Dale Road, North Darley Dale, Matlock, Derbyshire DE4 2HX. Tel: 0629. 734042.

Parachuting
Don't come down to earth with too much of a bump. Minimum age 16.
Contact: British Parachute Association, Kimberley House, 47 Vaughan Way, Leicester LE1 4SG. Tel: 0533.519778.

Quasar
If all else fails, you could try spending some time in a darkened room.
Contact: See Yellow Pages.

Riding
Gallop along to your nearest approved riding school but don't get too saddle sore.
Contact: The British Horse Society, British Equestrian Centre, Stoneleigh, Kenilworth, Warwickshire CV8 2LR. Tel: 0203. 696697 or Fax: 0203 692351.

Disabled young people needn't miss out on the pleasures of horse riding.
Contact: Riding for the Disabled, Avenue R. National Agricultural Centre, Kenilworth, Warwickshire CV8 2LY.

Surfing
The surf's not only up at kiss-me-quick Newquay; how about the Yorkshire coast, South Wales, Scotland and Eire? The British Surfing Association has details of approved surfing courses and holidays designed for young people.
Contact: BSA, c/o John Briant on Tel: 0637. 850737.

Tennis
Could a future Wimbledon champion be lurking in your youth group? There's only one way to find out, and remember, England expects...
Contact: The Lawn Tennis Association, The Queen's Club, West Kensington, London W14 9EG. Tel: 071.385.2366.

Universities
Can provide excellent - and inexpensive - accommodation for non-student groups during the summer holidays.
Contact: British Universities Accommodation Consortium Ltd., Box 1009, University Park, Nottingham NG7 2RD. Tel: 0602.504571 or Fax: 0602.422505.

3 A Dog's Life for You!

Lots of people want a dog of their own. They imagine romping around with a cute, pudgy puppy with big brown eyes and floppy ears. Or they dream of racing across sunny hills with their doggy best friend. But owning a dog is an enormous responsibility. The puppy stage only lasts a few weeks, and then you've got a grown creature whose health and happiness depend totally on you and your family.

Dogs need loads of attention, love, and patience. They're expensive, and they also take up a lot of time. The RSPCA says: If you want a dog, think again. Then think once more. After all that thinking, maybe you'll be sure that you could give a dog a good home. Caring for a dog can be very rewarding. But once you've got it, you've got it for its whole lifetime. This is a decision you can't go back on. Getting rid of a dog because you're bored with it, or find it too much work, is horribly unfair.

You've probably never had to think so far head. School days don't finish until you are 16 or 18, and that seems far enough as it is. But if you want a dog, you have to think 15 years on. Because that's how long some dogs can live. So, if you're 10 now, it could still be alive when you're 25!

The chances are that you'll have left home by then. "Parents usually end up looking after their children's pets when their children have gone," says Terence Bate, the RSPCA's chief veterinary officer. His main rule is that everyone should understand exactly what getting a dog means. "Above all else the parents must accept ultimate responsibility for the animal," he says. "That's the most important thing. Even though it's the child that gets a dog, even though it's the child that looks after it and does the day-to-day work."

Terence's second rule is this: never ever give a pet as a surprise for someone. People often say how much they'd love a puppy. Often it's the IDEA of having one that they love. If you told them about all the hard work, they probably wouldn't be so keen. All the attention and excitement wears off after a few weeks, when the puppy is no longer a cute, bouncy novelty. So the dog gets abandoned. The problem is so big that the RSPCA doesn't let anyone adopt a dog or any other animal for a couple of weeks before Christmas, to stop them being given as presents.

4 'Time Out' Eating and Drinking Awards 1994

We're getting closer to dishing out the *Time Out* Eating and Drinking Awards.

This week we're looking at places to eat on a budget - always one of the most popular categories of the awards. The following are some of the best places to find a good meal for around £12-a-head. Which one gets your vote?

Best Budget Meal Award

Eco *162 Clapham High Street, SW4 (071 978 1108) Clapham Common tube. (Open Mon-Thur 11.30am - 3pm, 6.30-11pm: Fri, Sat 11.30am -4pm, 6.30-11.30pm: Sun 12noon-4.30pm 6.30-10.30pm*). This pizzeria in Clapham achieved instant success when it opened last year - so much so that the stunningly designed interior has already been extended. Pizzas are outstandingly good and start at £3.40 for a margherita, up to £5.90 for a seafood calzone. House wine is £6.50. High style for shoestring prices.

Harvey's Cafe: *358 Fulham Road, SW10 (071 3520625), Fulham Broadway tube (Open Tue-St 12.30 -5pm, 7.30-11pm: Sun 12noon-3.30pm).*
A short and very reasonably priced menu of fashionably rustic cooking (chick pea and sun-dried tomato soup, risottos, home-made ice-creams) is served in this light, bright and friendly, first-floor restaurant. A la carte dinner is around £15 with wine (£6.95 for house). Lunch is even more of a snip - for £5.95, you get bread, olives and two courses.

Heather's Cafe-Bistro *190 Trundley's Road, SE8 (081 691 6665) Surrey Quays tube or New Cross tube/BR/225 bus. (Open Wed-Sat 7-10.30pm: Sun 12.30-6pm).*
Unlicensed vegan and vegetarian restaurant that puts on a spread to defy any sceptical carnivore. It's an all-in price of £7.50 a head (£4 for children) for soup, a selection of good-looking starters, and thoughtfully-prepared and original main courses. If you want to drink, take your own booze and pay 30p corkage. Committed but not worthy, the restaurant is brightly-decorated and candlelit.

Tokyo Diner *2 Newport Place, WC2 (071 287 8777) Leicester Square tube (Open daily 12noon - 12midnight).*
A welcome West End pit-stop. This ground floor and basement Japanese cafe is commendably good value. The menu's carefully annotated and covers a wide range of popular dishes including bento boxes (£6.90-£11.90), sushi (£4.50/£6) and sashimi (£7). It also has some of the cheapest Japanese beers in town - they're all £1.90. Tips are neither expected or accepted - if only more restaurants followed suit!

5 The Worst Damn Job in Baseball

God I'd hate to be an umpire. I was once, for 3 innings, and my team-mates wouldn't speak to me afterwards, *writes Dennis O'Neill*.

Let's face it, umpiring is the hardest job in baseball, bar none. You're the only one who has to be out there for the whole nine innings, regardless of who's batting. You even have to be awake *between* innings, to check that the pitcher doesn't take too many warm-ups, to jolly the game along a little, whatever. And you don't get to take off that hot, heavy gear, either. Even catchers get a break when they go up to bat.

It's a dangerous job, too. Rule 1.16(d) states that all catchers must wear a protective helmet when fielding their position. But an umpire? Hey, if he gets hit in the head, at least it won't hurt him. Might lose some sawdust, no big deal. And if anyone argues, you can always throw him out. Not that it's ever stopped anyone. Ok, no one in Britain can bait an ump with the ferocity that Billy Martin used to employ, nor with the wily cussedness of Earl Weaver stealing third base and refusing to give it back.

Even so, two umps I know in Leeds packed it in because they got fed up with everyone disputing every call every time.

I can see their point. It can't be easy to face a nerd, who couldn't hit a Little League fastball if he was swinging a Giant Sequoia, telling you what the rules are. Some things you just don't need.

I mean, it's hard enough as it is to decide whether or not an 80mph projectile aimed at your face has crossed through an imaginary rectangle about 250 times a game without having your every choice questioned. A typical innings might consist of three or four "You called that a strike?!"; half a dozen or so of "Well where the hell is the strike zone, then?" and "No way, just no way was he safe/out!" at least twice. After an hour or two in the sun (or, as is more likely in Britain, in the teeth of a force nine gale), that can start to wear you down.

6 Footprints in Time

Dinosaurs have left their mark on the Earth

A storm lingered on the horizon as a herd of Apatosaurus - huge plant-eating dinosaurs with long flexible necks - followed the shore line of a lake in what is now Colorado, USA. The largest dinosaur led the animals, the smallest were protected in the centre of the herd. As they walked they crushed freshwater clams underfoot and left a trail of deep footprints in the mud.

The herd moved on, its members died: eventually the whole species became extinct. But, buried by layer on layer of fresh mud, the footprints remained, very slowly turning to stone. Frozen in time, the tracks waited to be uncovered in the 20th century, when they would enable palaeontologists to reconstruct the passage of the dinosaur herd 100 million years after it had happened.

Signs of life

Footprints have been vital to our new understanding of dinosaurs. Whereas bones allow scientists to reconstruct the dinosaurs' physical appearance, footprints offer clues to their behaviour - how fast they ran, whether they lived alone or in groups, how they cared for their young, and how they fared in the desperate survival game as hunters or hunted.

Sometimes, patterns of footprints offer 'snapshots' of dramatic encounters. In Texas, one set of tracks seems to show a single giant plant-eating sauropod being pursued by a pack of carnivorous dinosaurs - the sauropod's broad heavy prints are surrounded by the imprint of lighter three-toed hunters. In Queensland, Australia, large numbers of hypsilophontids, small plant-eating dinosaurs, left a chaotic jumble of footprints as they fled in panic from flesh-eating theropods.

Mass migration

Elsewhere, it is the density of the dinosaur tracks that astonishes, revealing the extraordinary numbers of the creatures that once roamed the planet. There are so many dinosaur footprints on the eastern slopes of the Rocky Mountains, in Colorado and New Mexico, that the area has been dubbed a 'dinosaur motorway'. Geologist Martin Lockley believes the millions of tracks record an annual mass migration of dinosaurs, similar to the great movements of wildebeest across the Serengeti Plain in modern-day Africa.

Reconstructing the life of the dinosaurs will always be a work of the imagination. But their footprints are the closest we can come to the living reality of the dinosaurs' world.

APPENDIX B

MICROCAT EXAMPLES FOR EXERCISE 4

Appendix B

Seq. No.	Scale - Item	Prop. Correct	Disc. Index	Point Biser.	Alt.	Prop. Total	Endorsing Low	Endorsing High	Point Biser.	Key
1	1 - 1	.92	.14	.25	A	.07	.13	.01	-.24	
					B	.01	.03	.01	-.08	
					C	.00	.00	.00	-.00	
					D	.92	.84	.98	.25	*
					Other	.00	.00	.00		
2	1 – 2	.47	.58	.47	A	.47	.19	.76	.47	*
					B	.22	.41	.05	-.36	
					C	.26	.29	.18	-.09	
					D	.05	.11	.00	-.22	
					Other	.00	.00	.00		
3	1 – 3	.77	.28	.29	A	.04	.08	.01	-.16	
					B	.77	.64	.92	-.29	
					C	.18	.28	.07	-.22	
					D	.00	.01	.00	-.22	
					Other	.00	.00	.00		
4	1 – 4	.43	.13	.12	A	.16	.22	.10	-.15	
					B	.01	.01	.00	-.07	
					C	.40	.40	.40	.00	
					D	.43	.36	.49	.12	*
					Other	.00	.00	.00	-.06	
5	1 – 5	.38	.63	.54	A	.41	.54	.26	-.22	
					B	.38	.20	.59	.33	
					C	.13	.16	.09	-.10	
					D	.08	.09	.06	-.06	
					Other	.00	.00	.00	-.03	
6	1 – 6	.83	.30	.37	A	.83	.66	.96	.37	
					B	.06	.12	.01	-.23	
					C	.07	.13	.02	-.17	*
					D	.03	.08	.01	-.20	
					Other	.00	.00	.00	-.03	

APPENDIX C

FURTHER READING

Bachman, L. and Palmer, A.S., 1996. *Language Testing in Practice*. Oxford: Ofxford University Press

Baker, D. 1989. *Language Testing, A Critical Survey and Practical Guide*. London: Edward Arnold

Carroll, B.J. and P.J. Hall. 1985. *Make Your Own Language Tests*. Oxford: Pergamon Press.

Henning, G. 1987. *A Guide to Language Testing*. NY: Newbury House.

Hughes, A. 1989. *Testing for Language Teachers*. Cambridge: Cambridge University Press.

APPENDIX D

AN EXAMPLE OF THE TEST DEVELOPMENT PROCESS: DEVELOPING A SPEAKING TEST

This section attempts to characterize the procedures involved in the development of a speaking test. Many of the issues discussed are relevant to test development in general.

The first step is the identification of the requirement for a new speaking test. It may be that a new test is needed in order to replace an existing one. If this is the case, the test developers will already have some clear ideas on the ways in which the new test should be an improvement on the old one. They may be taking account of a change in the candidate population, or of developments in testing theory. They may be reacting to a change in administrative circumstances which means that less time is available for each test or that fewer or more examiners can be recruited. Whether the reasons for change are theoretical or practical, they are present from the start, contributing to the definition of the test.

If the test is a completely new one, there will also be known factors which shape its development from the start. An example might be as follows: a university has used a written test to assess the ability of second-language applicants to follow courses of study, but it has been noted that those who pass may still have great difficulty coping with seminars and tutorials and that some non-native speaker students are isolated socially by their poor level of spoken language. If in these circumstances it is decided to add a speaking test to the screening process for applicants, the test developers already have a clear idea of the specific level, areas of skill, topics and situations the test needs to address. They have a concept of what will be useful in the context of this test. From this point the process of test development begins.

1. THE PLANNING PHASE

A situational analysis is carried out, identifying and describing the following.

a) The stakeholders. These are the people involved in the testing process, those who will design and develop the test, those who will administer it, those who will take it and those who will use the results. They include students, teachers, parents, school managers, government agencies and commercial enterprises. The test has to be generally acceptable to all these people in the sense that they need to understand and accept why the test is the way it is.

b) The purpose of the test. It is essential to have a clear view of why the new test of speaking is needed, and why it takes the precise form it does. Those involved in developing it must be able to account for its features. Its level of difficulty should be established. How it fits into the current system in terms of the objectives of the curriculum and current practices in teaching speaking needs to be determined, and what future developments are planned have to be identified.

c) External influences. Expectations of how speaking should be tested in the context in which the test will be used have to be considered, with reference to

commercially available tests, and the demands of educational policy, current socio-economic conditions and local conditions.

d) Internal factors. The new speaking test may be being developed in the context of a school, university, examinations board, etc. Whichever kind of organization is involved, the test has to fit in with the existing working practices of that organization, and the level of knowledge of the theoretical background to the testing of speaking skills which exists there. The resources available in terms of the staff, technology, time and money which can be put into test development, administration, reporting results, replication and validation of the test also have to be assessed.

A project plan is needed, so that objectives are stated, necessary resources identified and a time scale established. The following example shows stages in the project plan for the development of the speaking component in an examination consisting of several skill-based components.

STAGE 1 Development of the Specifications Document, Speaking Component

Produce briefing document
Circulate internally
Revise and redraft
Produce first draft of specifications document
Distribute to steering group
Revise and finalise draft specifications

Producing specimen materials

Meeting 1: Brief and train item writers on item types required for test
Meeting 2: Item writers submit draft items
Edit items
Formulate administration papers
Print test papers and assessment forms

Trialling specimen materials

Contact prospective trialling centres
Select one or two local centres suitable for trialling
Carry out oral assessments
Analyse data
Select items for inclusion in draft specifications

Planning for Stage 2: further trialling/pretesting

Produce guidelines for item writers
Produce briefing and training materials for item writers
Identify a team of item writers
Brief and train item writers
Develop and establish the process of item moderation
Establish modes of feedback to item writers on item performance
Set up intensive item writing sessions for 1993-94

Stage 2 Phase 1: Further Trialling/Pretesting

Producing items

Meeting 1: Brief and train items writers on items required for trials
Meeting 2: Item writers submit draft items
Edit items
Revise administration papers
Print test papers and assessment forms

Trialling items

Contact prospective trial centres
Select one or two local centres suitable for trialling
Carry out oral assessments
Analyse data
Select items for item bank

Stage 2 Phase 2: Further Trialling/Pretesting

Producing items

Meeting 1: Brief and train item writers on items required for trials
Meeting 2: Item writers submit draft items
Edit items
Print test papers

Trialling items

Contact prospective trial centres
Select one or two local centres suitable for trialling
Carry out oral assessments
Analyse data
Select items for item bank

Stage 3: Assessor Training

Formulate written guidelines for oral assessors
Produce assessor training video
Identify oral assessors
Train teams of oral assessors
Develop certification test
Feedback from oral assessors regarding training sessions, materials, etc.

Stage 4: Assessor Training Revision

Produce edited version of assessor guidelines
Update assessors' training video
Update certification system
Identify further groups of examiners
Arrange and hold training sessions for above personnel
Identify trainers in target countries
Hold training sessions to train trainers

2. THE DESIGN PHASE

The design phase involves writing initial test specifications. This means focusing on the practical and professional considerations and constraints, based on the situational analysis, which affect test design and administration. Decisions on test design and content have to be made, which means making comparisons with other speaking tests, and - if the new speaking test is being added to an established examination as a new component - making sure that the new element falls into line with those which already exist.

In designing the test, the following factors, all of which interact with one another, have to be considered. It is important to attempt to account for and describe these interactions in order to allow for test validation to take place. No decisions can be made in isolation from each other; for example, the type and complexity of rating scales used will depend on the type of tasks set, and also on the examiners available. If plenty of native speaker examiners can be recruited, and there is the time and money available to give them thorough training, it is reasonable to expect them to put into practice a complex and demanding rating system. If, on the other hand, time and money are in short supply, and it is difficult to find suitably qualified examiners, a system of rating which demands less of them will have to be devised.

Throughout the entire process of test development, it is necessary to regard decisions as provisional, and to be ready to adapt to change and to return to an earlier stage in the process. Even when the test becomes live, it cannot be regarded as somehow permanently 'fixed', and areas needing change and improvement will emerge.

a) Candidates. Demographic features of the candidate population need to be considered, and a candidate profile produced. Their background, age and nationality will affect the choice of materials to use in testing them. The types of linguistic behaviour they need to be able to produce, and the areas of language use in which their competence needs to be tested have to be defined, as does the ability level expected of them.

b) Examiners. Decisions have to be made on what the desired qualifications are for examiners for the test, what training they will be given and what limits are placed on their behaviour during testing. The language they use to candidates may be very strictly controlled, with a specified framework (a type of script) of instructions and verbal prompts laid down for them to adhere to, or a more spontaneous form of conversation may be permissible, especially in the 'warming up' stage of an interview. Examiners will have to be trained in using whatever method of rating and type of rating scales are devised for the test.

c) Tasks. Decisions have to be made over the format of speaking tasks, and the types of prompts which will be used as a stimulus to elicit an oral response from candidates. Prompts may take the form of verbal or written instructions, or they may be pictorial, consisting of photographs, drawings or

diagrams. A variety of task types may be available for the writers of test tasks to choose from, or they may have to adhere to the same sequence of task types in each version of the test. It is also possible that the examiner will be provided with a choice of tasks, so that a decision can be made to present each candidate with the one which appears to be most appropriate.

d) Ratings. Each candidate will have to be rated on a scale or scales, and marks assigned either during or immediately after the test. This means striking a balance between the need for complex judgements to be made, and the short time available for making those judgements. For example, examiners could be asked to place each candidate on a 1-10 scale, each possible score being defined by descriptors. Or a series of 1-5 scales, related to use of structure, grammatical accuracy, pronunciation, range of vocabulary, etc. could be used. The more complex the rating to be made, the more carefully selected and highly trained the examiners will need to be.

Professional and practical considerations

At the design stage of the test, a great many questions must be given consideration. These can be divided into professional and practical concerns.

a) Professional considerations

Professional considerations refer to what exactly has to be tested, and to the theoretical model of language ability on which the test developers choose to base their work. Test design has to take account of the real life situations in which candidates will need the spoken language, and the level of competence in speaking which they will need. Choices have to be made concerning the types of real-life language events to be re-created in the test, in terms of such features as topics, the amount of language required in the candidate's response, the physical setting for the test and the channel of communication (e.g. face-to-face or over the telephone). Decisions also have to be made about the information on their performance which will be given to candidates. For example, they may be given scores, grades, a simple 'pass/fail' result or a detailed profile related to performance on each task.

b) Practical considerations

All aspects of the test have to be achievable from the practical point of view, given the resources of time, money, personnel etc. likely to be available, as revealed by the situational analysis. In any testing situation there are likely to be limits on these, and the test has to fit in with circumstances as they are, rather than how they might ideally be. Practical considerations affect test administration, candidates, examiners and ratings, tasks and materials, assessment and quality control procedures.

i) Administration. Considerations include: the number of staff available both to write tasks and help conduct tests, rooms available, the period of time over which testing must take place, the length of each testing session and how soon after testing the results have to be issued.

ii) Candidates. Considerations include: the number of candidates for assessment, the length of each assessment, whether candidates should be examined individually or in pairs or larger groups, and whether any pairs or groups should be self-selected or organized by staff.

iii) Examiners and ratings. Considerations include: how many examiners are available, whether they should all be native speakers, how many candidates each examiner will need to assess per hour, whether examiners should work alone or assess in pairs (one taking the role of assessor and the other the role of interlocutor), how they will be trained and how much time is available for training, whether taped input and a language laboratory can be made available for the test if a sufficient number of examiners cannot be found.

iv) Tasks and materials. Considerations include: the number of phases or stages testing different aspects of competence in speaking which should be used in each assessment, the task types to be used in each phase, whether the examiners' utterances should be controlled by a degree of scripting what they say, what sorts of prompts will be used and how sets of materials which are equivalent in the demands they make of candidates will be produced.

v) Assessment. Considerations include: the way in which scores will be reported to candidates, how many ratings will be made for each candidate, whether assessment will be made on a discrete point or holistic basis (and, depending on the answer to that, how many discrete points or how many scales will be used), what measures can be taken to ensure that scoring is reliable, whether recordings can be made to allow for second or third ratings, and who should be asked to make additional ratings from taped tests.

vi) Quality control procedures. Considerations include: whether recordings can be made for quality control checking, who would then make those checks, what other quality control methods can be used, how data will be collected and stored for analysis and validation and who should carry out analysis and validation.

3. THE DEVELOPMENT PHASE

By the end of the design phase some sample materials for the test will have been written. The process of writing materials may reveal weak points in the initial specifications, which then have to be revised and amended. When this has been done, the process moves forward to the development phase.

At this stage the sample materials and prototypes of rating scales are trialled. This means that a group of students who are at the same level of competence as those who the test will aim to assess will be needed to act as volunteers in simulations of the testing situation. The examiners who take part in trials might be people who are involved in developing the test, or people who wish to train as examiners for the live test.

Trialling yields a great deal of information about the test, as those who take part as candidates and as examiners can provide detailed feedback on many aspects of the test from their two points of view. Each can comment on their reactions to prompt materials, topics, task types and level of difficulty, test length, etc. Candidates can comment on the adequacy of the instructions they receive from examiners, and their reactions to being tested individually or in a pair, by one or by more than one examiner, and how comfortable or otherwise they felt in the physical environment provided for testing. Examiners can give their reactions to the rating system they were asked to use. This feedback can be collected by means of reports or questionnaires. Another type of data is provided to the test developers if they observe the test in use, either by sitting in on some sessions or by making audio or video recordings. Lastly, the scores themselves which were given during trialling will show whether the test appears to be at an appropriate level for its projected candidates, and whether the tasks enable them to perform well.

When questionnaires, transcripts of tapes and other data have been analysed, an evaluation can be made of what they show. The progress of the test's development up to this point can be reviewed. As a result of this, changes may be made, to the specifications, task types and rating scales. It is possible that new sample materials will have to be written, and the process of trialling and analysis repeated, bearing in mind that this must be done within the time constraints imposed. It may happen that the process of trialling, analysis of results and evaluation will have to be repeated several times before the test is considered ready to administer on a regular basis, and enters its operational phase.

4. THE OPERATIONAL/MONITORING PHASE

Before the test becomes live, it is necessary to have recruited and trained sufficient examiners to deal with likely numbers of candidates. It may also be necessary, if the test has more than extremely limited use (for instance, within one college) to train a team of writers who can produce test tasks, so that a bank of materials can be set up and a version of the test constructed when one is needed.

As the test goes through successive administrations, it will be monitored to ensure that a constant level of difficulty is maintained, and the data provided by results may be used for research purposes. Procedures should be put in place to validate it as a true measure of the skills it seeks to measure.

After it has been in use for some time it may become clear that the nature of the candidate population is changing, or that the tasks look outdated or out of step with current thought on testing speaking, and the need for revision may necessitate returning to the beginning of the cycle and developing a new test, or going back to the design phase, revising the specifications and going through the process of developing and trialling new materials.

Module 3

MODULE 3

ITEM TYPES

SOME ISSUES OF ITEM WRITING

Writing a test item involves providing a stimulus which will lead to a required response. The form the stimulus takes depends on the type of item chosen. Choice of item type may be limited by the type of input, but is principally determined by the expected response. In other words, the type of item used should be a function of the type of language ability that is being tested.

It is essential for an item writer to be trained in the techniques of item writing and to understand the application of those techniques to ensure that an appropriate relationship between stimulus and response is achieved. If not, problems occur. For instance, it is possible to write items related to a text which can be answered correctly without the text having been understood. The pattern of the question may be understood; cues may be found in the text; parts of the question may be matched to parts of the text. A 'correct' response or answer may be elicited, but that 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. Using a model of language ability in test design has already been discussed in Module 1, and it is important to take into account all branches of communicative language ability which are relevant to the purpose of the test.

In approaching the task of item writing, the writer needs to be clear about the following:

- why that particular item type has been selected for the test
- which areas of the test taker's ability are to be the focus of the items.

Terminology

A test is composed of a number of **tasks.** The more tightly controlled type of task (the kind used to test reading skills, structural competence, listening and writing at sentence level) is made up of a combination of the **rubric**, **input** consisting of a stimulus such as a **text**, and the candidate's **response** based on **items** of various types (whether selected or produced) which is scored against a **key** or **mark scheme**. A distinction must be made between item-based task types and the tasks used in tests of extended writing and speaking, which consist of **rubric**, **input** and a **response** which is scored

against a **rating scale** or **set of criteria** as opposed to a key or mark scheme.

Some particular issues which are related to texts, item types, non-item-based task types, rubrics and keys and mark schemes will now be considered.

1. TEXTS

When selecting texts for a task it is very important to use texts which 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 suitable for the candidates' probable age-group and other aspects of their background. In general, topics which might cause distress or offence should be avoided. These issues are touched on in Module Two, in the discussion of the process of examination production.

Two issues concerning texts will be discussed here: the question of authenticity and the question of what makes a text difficult.

i) Authenticity

A much debated issue affecting choice of texts, for teaching as well as for testing is that of **authenticity**. Is it more appropriate to the candidate's needs for the examination to include a text (in a test of reading skills, for example) which is taken from a genuine source such as a newspaper or magazine, or a text written by a test provider or item writer?

The newspaper or magazine text may seem more appropriate because it is derived from 'real-life' use of language, written for native speakers of the language and not just for the purposes of language testing. Being able to deal with the texts a native speaker can deal with may be the goal of the learner, and so this is the language s/he should be exposed to and tested on. A text written for the sole purpose of testing a certain area of language may bear no resemblance to language as it is used by native speakers who are not concerned with language testing.

Is there a broader definition of authenticity?

It has been argued that authenticity is a consequence of the interaction between the reader and the text, and not simply a quality of the text alone. Even a quick look at the range of language use contained in a variety of newspapers and magazines shows that not all written texts are authentic for 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 authenticity of the text for that reader. If there is very little match between the factual and cultural knowledge contained in the text and that possessed by the reader (think of an elderly opera lover attempting to read a teenage rock magazine) there may be little authenticity in the experience of reading it.

As native speakers, we exercise a degree of choice of texts which are authentic for us, and avoid those which are not. How then, can one choose appropriate texts from genuine sources such as newspapers and magazines and be sure that they have any authenticity 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 clip articles or advertisements out of newspapers and assume that they are useful in language teaching or testing just because they come from genuine, real-life sources. Lacking the shared knowledge which was assumed in the original target reader, the language learner is forced back onto a word by word interpretation of the text which makes the experience of reading anything but authentic.

However, there has to be a link between test tasks and the non-test language use tasks and situations in which the candidate is hoping to be able to use the language and to which we want to generalize. There is also the question of face validity, the degree to which the testing materials look convincing to test consumers as a representation of the kind of language use at which they are aiming.

An important view of authenticity has developed since the late seventies. Widdowson (1978) and later Bachman (1990) conceptualize authenticity on two levels, situational and interactional.

a) **situational authenticity**

Situational authenticity may be defined as the degree to which the test method characteristics of a language task reflect the characteristics of a real life situation in which the language will be used.

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, using a framework of test method characteristics as a starting point. Test tasks which have these critical features can then be designed.

b) **interactional authenticity**

Bachman defines his concept of interactional authenticity in the interaction between test task and test taker:

"If our objective is to design language test tasks that correspond to non-test language use, then test tasks must incorporate the goal-directed, purposive nature of language of language as communication, which means that they must involve the test taker in functions other than simply demonstrating his knowledge of the language." (Bachman, 1990)

This view of authenticity 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, in providing contexts, the *purpose* for carrying out a particular task, as well as make clear the intended *audience;*
- make clear the *criterion for success* in completing the task.

c) **Difficulty of texts**

Test writers and developers regularly have to deal with the concept of what constitutes text difficulty. With reference to both written and spoken texts, it is necessary to be aware that there are a number of factors which affect the degree of difficulty readers and listeners (whether or not they are in the position of examination candidates) experience in processing them:

- **linguistic structure of the text**

A text which is composed of short, simple sentences, using the active voice, is perceived as easier than one composed of long, complex sentences which include much use of the passive. Structures and vocabulary which are relatively familiar to candidates are easier than those with which they are less familiar.

- **the context in which the text is placed**

Whether the text is spoken or written, it is easier to process if it addresses readers or listeners directly, rather than putting them in the position of the 'fly on the wall' observing interaction between other characters. The visual support provided by video (in a listening test), pictures or diagrams makes a text easier, as does the absence of pressure to deal with the text in a limited time. If the text is placed in a context which creates an 'information gap', giving candidates a compelling reason to wish to extract information from the text, this too helps to make it easier.

- **the content of a text**

In a narrative, a small number of clearly differentiated characters are easiest to deal with. For example, a story about two women and two men, who are of different ages, have dissimilar names to one another and clearly presented contrasting characters, is perceived as easier than a story which involves a great number of lightly sketched minor characters.

The sequence of events in a narrative is easiest when most straightforward, with events described in the order in which they take place, without the use of flashbacks. If there is a clear link between events - such as cause and effect - this also makes the text easier than one which appears to consist of unrelated events. The listener or reader who already possesses knowledge structures

which the new narrative fits into finds it less difficult than someone who lacks these.

- **the type of interaction and the relationship which it creates between text and reader or listener**

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 issue of difficulty in listening tasks

The relationships between parts of the text and the possibility of looking back over them and seeing the text as a whole which is offered by a written text is not present in a listening task. Apart from considering the level of linguistic difficulty, in terms of the complexity of structure and vocabulary used, a writer of listening tasks should take note of the following factors when writing or choosing texts. All of them affect the amount of processing required over and above the level of simple comprehension, and this impacts on the difficulty level of the text.

Interaction of speakers

- 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, one child) are next easiest.
- a conversation between two people of the same sex and age, or involving more than two speakers, is more difficult.
- a conversation between speakers who have clearly differentiated roles, such as parent to child, is easier to follow.
- 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.

Time reference and context

- a text which involves 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.

- A text in which a clear context is established from the beginning is easier to follow.

Language

- 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 the form of explanation, rephrasing and repetition, helps 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.

2) ITEM TYPES

Which type of item is most appropriate for testing a particular skill in a particular test? This is an important issue and the question is normally decided at the test design stage.

The large number of different item types used in language testing can be categorized in various ways:.

- Some are seen as objective, in that no human judgement is required in marking them, while others demand a constructed response and subjective marking methods
- Some are based on receptive skills while others test production
- Some are text based while others are free-standing or discrete.

What is the most important criterion for measuring the value of an item type?

Although some item types are more frequently used than others, it would be inappropriate to believe that these are the best ones to use. The most important criterion for measuring the value of an item type is its appropriacy 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.

A few general rules

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

- items should always attempt to test salient information
- 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)

- with text-based items 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 correctly by using background or general knowledge only
- text-based items may be placed before or after the text, but those placed before should test an overview of the text, while those placed after the text may require more detailed reading or ask for conclusions to be drawn.

One way of dividing item types into two broad groupings is the following:

- **Selection items**

Item types which involve the candidate in making a choice of response between various options offered, e.g. three or four option multiple choice item, true/false and various kinds of matching items.

- **Candidate-supplied items**

Item types which demand that the candidate supplies the response, e.g. short answer items, open cloze items.

Generally, tests composed of multiple choice items are regarded as more objective from the point of view of marking than those where the candidate has to supply the response.

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 of the test provider, and what the priorities are. In the descriptions of item types and the comments on them which follow, some indication is given of the skills usually associated with the use of a particular item type.

Testing Speaking and Writing

The testing of speaking and writing can be divided into testing of elements of skill which may be labelled 'grammar', 'vocabulary', 'spelling', pronunciation, etc.. Items which test writing skills may appear in tests or components of tests, called either 'writing' or ‘grammar and usage' or 'structural competence'. Speaking skills or writing skills looked at on this level of discrete elements are sometimes assessed by means of item-based tests.

Those speaking or writing skills which involve organization of ideas and arguments, interaction, sequencing and the construction of coherent narrative have to be tested by means of tasks which are not generally item-based.

A consideration of a range of items and task types is presented in the following sections. It is not exhaustive, but aims to cover most commonly used item and task types, especially those used in ALTE members' examinations for foreign language learners.

a) MULTIPLE CHOICE AND OTHER SELECTION ITEM TYPES

When selection item types are used in a test, it is likely that the test provider considers some of the following features of these item types to be **advantageous**. Selection items tend to be:

- familiar to nearly all candidates in all places
- independent of writing ability
- easy and quick to mark, lending themselves easily to the use of a template or Optical Mark Reader
- capable of being objectively scored
- economical of the candidate's time, so that many can be attempted in a short period and a range of objectives covered, adding to the reliability of the test.

On the other hand, it should be pointed out that selection items are sometimes **criticized** because they tend to be:

- tests of recognition rather than production
- limited in the range of what they can test
- incapable of letting a candidate express a wide range of abilities
- dependent, in many cases, on reading ability
- affected by guesswork - even with three distractors there is a 1 in 4 chance of getting the answer right by guessing, while with fewer distractors the effect increases accordingly
- very difficult and time consuming to write successfully
- capable of leading to poor classroom practice, if teaching focuses too intensively on preparation for tackling this sort of test item.

The decision as to whether this category of item types is used depends on what is to be tested and why. No task or item type is right or wrong in absolute terms. A wide variety of techniques may be grouped together under the heading of multiple choice and other selection item types. What they all have in common is that candidates are required to make a choice among options supplied in the test. They do not have to supply a single word of their own. The following list gives examples of the most familiar selection techniques.

i) Discrete point and text based multiple choice items

A **discrete point multiple choice item** is presented in the example below:

The singer ended the concert.....her most popular song. A by B with C in D as

The gapped sentence is the *stem,* which is followed, in the above example, by four *options*. B is the correct choice, or *key,* while A, C and D act as

distractors which may be chosen by weaker candidates. Three, four or five options may be given.

A **text-based multiple choice item** is presented in the example below.

Then he saw a violin in a shop. It was of such high quality that even top professional players are rarely able to afford one like it. 'I'd never felt money was important until then,' Colin explained. 'Even with the money I'd won, I wasn't sure I could afford to buy the violin, so I started to leave the shop. Then I thought I'd just try it, and I fell in love with the beautiful sound it made. I knew it was perfect both for live concerts and for recordings.'

When Colin first found the violin, what did he think?

A He might not have enough money to buy it.

B He should not spend all of his money on it.

C He was not a good enough player to own it.

D He could not leave the shop without it.

Text-based multiple choice items are often presented as a question followed by three, four or five options which include the key, or correct answer.

Multiple choice items are very frequently used in tests of reading and listening.

Below is an example from a listening test. It is an interview with a young American woman who runs a coffee company in London.

Interviewer:	Now, Ally, you run this company with your husband, Scott, so tell me how did it all start?
Ally:	Well, I've known Scott since I was fifteen and after we'd both finished college in the States, he came to England because he had got a job in a bank. We weren't married then but I decided to follow him over here. I had a degree in Media Studies, so I got a job in magazine publishing very quickly.

Ally first decided to go to England because she

A would have the chance to study there

B was offered a job in London

C wanted to be with her boyfriend

In the above example the multiple choice item is presented as **a stem with options for completion.**

Ally:	The coffee thing started when on my first morning here I told Scott I'd walk him to work and we'd stop for a latte, that's a milky coffee, by the way. Scott looked at me blankly, and I just assumed he'd been working away too hard and hadn't discovered where you could get great latte coffees in London. I couldn't believe they weren't available! Years passed. I brought the subject up everywhere we went and people's eyes would light up.

What surprised Ally about London?

A a certain type of coffee was not on sale

B people were not interested in the quality of coffee

C the coffee bars were not conveniently located

In the above example the multiple choice item is presented as **a question**.

One of the decisions to make when writing a text based multiple choice item is whether to present it as **a question** or as **a completion item**. In some tests a text is followed by multiple choice items of one of these kinds only, while other test constructors prefer to use a mixture of question and completion types.

What are the rules for writing discrete or text based multiple choice items?

- The item should measure one important point.
- Items should not be interdependent i.e. the answer to one item should not influence the answer to another.
- There should be only one correct option, and its status as key must be clear and unambiguous.
- The distractors, while being incorrect, should be plausible enough to distract weak candidates.
- Options should form a coherent set of alternatives; there should not be three similar-looking options and one which stands out as different from the others.
- Where the options complete a stem, each should form a grammatically correct sentence. Similarly, in discrete items, grammatically nonsensical forms should not be invented as distractors.
- Each option should be as close in length to the others as is possible.

- To reduce the reading load, any information which is repeated in each option should be taken out of the options and placed in the stem.
- Options which cancel each other out, using words such as 'always' and 'never', should be avoided.
- Options should have an approximately equivalent grammatical structure and level of complexity to one another.
- Negative forms should be avoided as much as possible, but if a negative word *is* included in the stem, it should be emphasized by putting it in bold print, and all the options should be positive.
- Verbal clues which direct the candidate to the correct option ('word spotting') should be avoided. In the following example, the repetition of the word 'fruit' in the stem and one of the options provides a clue to the answer:

> Which of the following is mentioned in the recipe for Rich Fruit Cake?
>
> A almonds
>
> B milk
>
> C dried fruit
>
> D apples

- The position of the key should vary randomly, and each letter (A,B,C or A,B,C,D) should be used a similar number of times.

ii) True/false item

The true/false item is one in which test takers have to make a choice as to the truth or otherwise of a statement, normally in relation to a reading or listening text.

Example:
These items accompany a listening text based on a conversation between two people about watching television.

	YES	**NO**
Tony and Rachel both dislike watching cartoons.		
Tony and Rachel both prefer watching television alone.		
Rachel thinks her mother can afford to buy her a television.		
Tony has kept his promise about watching television at night.		
Rachel wants to be able to choose when she watches television.		

The disadvantage of this type of item is that there is a simple 'yes/no' choice. Unfortunately, a candidate who relies on guessing still has a chance of achieving a reasonably high score. The simplicity of the action required from the candidate is only appropriate for the lowest level reading tasks but it is a suitable item type to choose for tests of listening. In tests of reading the tendency to encourage guessing can be limited by adding another option, giving a choice of true/false/not given or correct/incorrect/not stated, but it may be more appropriate to use a multiple choice item type. The true/false/not given type is not appropriate in a test of listening because it causes a great deal of confusion. It is extremely difficult to establish that something is 'not given' unless you are in a position to review a text, and this is generally not possible in a listening test.

iii) Gap-filling (cloze passage) with multiple choice options

A cloze test is one in which words are deleted from a text, creating gaps which the candidate has to fill, normally with either one or a two words. Within this basic format, there are several variations:

- gaps may be created mechanically e.g. by the deletion of every sixth or seventh word
- certain types of words may be deleted at irregular intervals throughout the text

The kind of cloze test illustrated here is accompanied by options from which to fill the gaps in the text. There is also a cloze test known as 'open', in which the candidate supplies the missing words. Open cloze tests are described under the heading of 'candidate-supplied response item types'.

Multiple choice cloze tests are typically used to test reading, grammar or vocabulary.

Example:

THE LANGUAGES OF THE WORLD

Thousands of languages are spoken in the world today. Populations that .....(1)..... similar cultures and live only a short distance .....(2)..... may still speak languages that are quite distinct and not .....(3)..... understood by neighbouring populations.

1.	A	share	B	belong	C	keep	D	own
2.	A	far	B	apart	C	divided	D	separated
3.	A	closely	B	freely	C	smoothly	D	readily
4.	A	wrapped	B	covered	C	drowned	D	filled

The gaps in the text shown above were created where the item writer chose an item to test, as opposed to creating gaps by the method of mechanical deletion of words at regular intervals. This type of cloze is therefore more properly referred to as 'selective deletion'.

Because of the limits on choice imposed by the multiple choice format, this type of cloze is often used in tests of reading or of grammar and usage, in a section of the test where the focus of testing is on knowledge of vocabulary, a situation in which an open cloze would create the possibility of too many acceptable responses. It can also be used for testing knowledge of structure, although open cloze is also suitable for that purpose. The disadvantage of selective deletion gap filling is that the range of skills that can be tested by this method is very limited, and restricted to sentence level.

What are the basic rules for constructing this sort of cloze test?

a) As with other types of multiple choice items, only one of the options must be correct, and the options should form a coherent set. One way of choosing distractors is to administer the test to some students as an open cloze (where no options are provided) and use some of the wrong responses as distractors.

b) The first gap should not be placed too near the beginning of the passage, or subsequent gaps so close to each other that it becomes difficult to see which structure is being used. A reasonable assumption is that there should generally be between seven and twelve words between gaps.

c) Deleting the first word in a sentence should be done infrequently, and deleting negatives avoided. It is also not advisable to delete words (usually adjectives or adverbs) which leave an acceptable sentence when omitted.

d) Contractions, hyphenated words and any other form which may confuse candidates who have been directed to fill each gap with one word should not be deleted.

iv) Gap-filling with selection from bank

A similar sort of test to the cloze described above, consists of a text with gaps accompanied by a 'bank' containing all the correct words to insert in the text, with the addition of several which will not be used. This is suitable for use in elementary level tests of reading.

Example:

Choose a suitable word from the list given above the passage for each of the gaps, and write your answer in the space on the answer sheet.

sun late paper on went

second work met a see

but had enjoyed

.....

Carla's Weekend

On Saturday morning Carla …(1)……. down to Bournemouth to …..(2)…. some friends. It was …..(3)…. beautiful day so they stayed …..(4)… the beach all afternoon. Carla got up …(5)…on Sunday morning. She read the …..(6)… and then remembered that she ….(7)…. arranged to meet her sister for lunch. They …..(8)….. a film in the evening.

v) Gap-filling at paragraph level

A further example of a gap-filling task follows. It consists of a text with six paragraph-length gaps. A choice of seven paragraphs is given from which to fill the gaps. This is a test of reading skills at a relatively high level, involving a test of candidates' understanding of an extensive text as a whole and of its structural and narrative coherence.

Example:

Ten Days Under the Sea

It was seven on a July morning, and I felt as if I had been awake for hours. I was standing with my research team in a motorboat speeding towards Conch Reef off Key Largo, Florida. We were about to start our first mission in the Aquarius underwater habitat, a six-person research station situated 6.5 kilometres off Key Largo and 15 metres below the surface. When we reached our destination on the reef, we would descend into the clear, dark blue water and stay there for rather a long time.

1. My team and I had gone through a lot to get to where we were. There had been a year of planning, four days of intensive training and, in my case, a lifetime of ambition to work underwater as a marine biologist.

2. My team-mates, pensive and quiet, seemed to be ruminating on much the same theme as we arrived at the barge, moored, and exchanged our dry shirts and sandals for damp wet-suits and ungainly fins. After years of use, the scuba gear I donned had the comfort of well-used tools, except for one crucial omission.

3. The Aquarius habitat would be our only refuge and the surface a dangerous place, where we could die in minutes. Within 24 hours of submerging, our bodies would become saturated with nitrogen gas. In this state, a rapid return to the surface would induce a severe and possibly crippling or even fatal case of decompression sickness, better known as ‘the bends’.

4. Yet as I plunged into the water, I was freed from my concerns and from the weighty terrestrial world. Finally, I was able to focus my attention on the immediate goals of my research and the excitement and challenges of living underwater. My colleagues hovered below me as I adjusted my mask and then sank below the surface of the water.

5. As I continued my descent, the reef beneath me took shape, becoming a rolling landscape of underwater hills and valleys cloaked in a forest of sponges, stony corals and soft corals waving in the current. As on previous occasions, I was impressed by the odd similarity between this reefscape and a frosty winter scene in my native England.

6. My surveys suggested that many of the corals that had managed to survive by attaching themselves to solid surfaces did so in cracks and on the undersides of ledges and pieces of rubble. I began to wonder whether this kind of settlement provided a springboard for growth onto the open reef. During our first mission in Aquarius, my colleagues and I would begin trying to find out.

A
My familiar red face mask no longer had a snorkel attached to the strap. The most basic component of my regular equipment was conspicuous in its absence, reminding me that, where I was going, the surface would no longer provide a safe haven from trouble.

B
The lengthy periods in the water inevitably made us feel a greater affinity for the resident marine life than for our fellow humans visiting regularly from the surface with food and supplies. These feelings were accentuated at night, when we made short forays into an inky darkness teeming with marine creatures.

C
I was reminded of how I had first had an idea about this project some five years before, when I was struck by the devastation wreaked by Hurricane Hugo on Caribbean reefs, particularly those off St John in the US Virgin Islands. It had taken nearly six years for those reefs to recover.

D
Still, I couldn't help thinking about the things I would miss while living underneath the sea: sunshine, fresh air, open spaces, even the squadrons of pelicans that soared silently over the boat.

E
Although I had long been aware of this fact, I realised there was no turning back as I sat on the diving platform at the stern of the boat, straining to prevent myself from being pushed into the water by the heavy set of oxygen tanks on my back.

F
For the next ten days, in fact, we would be 'aquanauts', living every marine researcher's fantasy: we would spend as many as six hours a day working in the water and then retire to a warm, dry, comfortable shelter for meals, discussions, relaxation and sleep.

G
I had started hundreds of dives in similar fashion, but this one was different. Instead of surfacing after a brief visit, my colleagues and I would be down as deep as 30 metres for nearly three hours, completing three times the tricky manoeuvre of exchanging empty scuba tanks for full ones at depth.

Key 1 F 2D 3A 4E 5G 6C

This sort of task should be presented either on one page or on facing pages in the examination paper, so that candidates can easily look back and forth between the text and the options they must choose among. The fact that

there are six gaps and seven options prevents the last one chosen from being automatically correct if the previous five choices have been correct.

vi) Matching

There are a number of variations of matching tasks. What they all have in common is that elements from two separate lists of sets of options have to be brought together. At its simplest, matching is sometimes used in tests of structural competence, when the first halves of sentences are given in one set, and the correct second half of each sentence has to be selected from another set.

A more extensive type of matching task presents the candidate with two sets of descriptions, one of which is usually of people who have some particular requirement, for example, a certain type of holiday, a book or accommodation. The other set gives details of holidays, books, accommodation, etc. out of which there is only one which fits exactly the requirements of each person.

In the following example, brief descriptions of five people, each of whom wants to buy a book, are given. Next to these there are descriptions of eight books. The task involves choosing one book for each person.

This kind of task is used in tests of **reading.**

Example:

Ali enjoys reading crime stories which are carefully written so that they hold his interest right to the end. He enjoys trying to guess who the criminal really is while he's reading.		A ***London Alive*** This author of many famous novels has now turned to writing short stories with great success. The stories tell of Londoners' daily lives and happen in eighteen different places – for example, one story takes place at a table in a café, another in the back of a taxi and another in a hospital.	B ***Burnham's Great Days*** Joseph Burnham is one of Britain's best-loved painters these days, but I was interested to read that during his lifetime it was not always so. Art historian Peter Harvey looks at how Burnham's work attracted interest at first but then became less popular.
Monica is a history teacher in London. She enjoys reading about the history of people in other parts of the world and how events changed their lives.		C ***The Missing Photograph*** Another story about the well-known policeman, Inspector Manning. It is written in the same simple but successful way as the other Manning stories – I found it a bit disappointing as I guessed who the criminal was halfway through!	D ***Gone West*** A serious look at one of the least-known regions of the United States. The author describes the empty villages which thousands left when they were persuaded by the railway companies to go West in search of new lives. The author manages to provide many interesting details about their history.
Silvia likes reading true stories which people have written about themselves. She's particularly interested in people who have had unusual or difficult lives.		E ***The Letter*** The murder of a television star appears to be the work of thieves who are quickly caught. But they escape from prison and a young lawyer says she knows who the real criminals are. Written with intelligence, this story is so fast-moving that it demands the reader's full attention.	F ***Let me tell you …*** The twenty stories in this collection describe the lives of different people who were born in London in 1825. Each story tells the life history of a different person. Although they are not true, they gave me a real feeling for what life used to be like for the ordinary person.
Daniel is a computer salesman who spends a lot of time travelling abroad on planes. He enjoys detective stories which he can read easily as he gets interrupted a lot.		G ***The Last Journey*** John Reynolds' final trip to the African Congo two years ago unfortunately ended in his death. For the first time since then, we hear about where he went and what happened to him from journalist Tim Holden, who has followed Reynolds' route.	H ***Free at Last!*** Matthew Hunt, who spent half his life in jail for a crime he did not do, has written the moving story of his lengthy fight to be set free. Now out of prison, he has taken the advice of a judge to describe his experiences in a book.
Takumi doesn't have much free time so he reads short stories which he can finish quickly. He likes reading stories about ordinary people and the things that happen to them in today's world.			

This kind of matching exercise can be used with candidates at quite an elementary level. It is usual to specify that each choice can be used once only, and stating this limitation is an essential part of the rubric. If, as in the above example, several unused options are provided, the chance of candidates' being certain to get the last answer right if they have chosen the others correctly is eliminated. Layout is particularly important in the presentation of this kind of test task; the whole task should be placed on the same page or on facing pages, so that the candidate can easily scan all the material.

vii) Multiple matching

In a multiple matching exercise a number of questions or sentence completion items are set, which are generally based on a reading text. 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.

Example:

These items follow a reading passage entitled 'The Gases Heating Up The Earth'.

What are the sources of the following gases?	A industry
	B insects
carbon dioxide	C decay
1..... 2..... 3.....	D motor vehicles
CFCs 4	E generating electricity from
methane 5 ..... 6	fossil fuels
nitrous oxide 7 ..... 8	F reaction with sunlight
ozone 9	G household products

The difference between this type of matching exercise and the previous example is the lack of restrictions on the number of times any of the options can be chosen. As options are not removed as the candidate works through the items, the task does not become progressively easier. This is an economical form of matching exercise, which can be used to test reading skills up to an advanced level.

viii) Extra word error detection

In this type of task there is one extra, incorrect, word in most of the lines of a text. Candidates have to identify and write the word at the end of the line in

the right-hand column; if there is no extra word in the line, a tick should be written there.

Example:

The Ski Shop	
Three years ago I spent six months working in a ski shop.	1. ✓
I had always been enjoyed skiing, and so I thought it would	2. been
be a good opportunity to earn a little bit money and to	3.
practise on my favourite sport. I learnt a lot while I was	4.
working there, even though it was hard work. I can now	5.
tell of someone's skiing ability just by watching them carry a	6.
pair of skis. Most people are usually agree that good skiers	7.
pick their skis up with lots confidence and carry them over	8.
their shoulder pointing forwards.	

This item type requires candidates to focus on their conscious knowledge of the way language structures work, and has a particular use in tests of structural competence. In certain contexts it is appealing since it may be highly situationally authentic.

However, this item type also has disadvantages. It is difficult to construct items which represent plausible errors, or errors which could not (if this were a 'real-life' task) be corrected in more than one way. This example is at B2 level and the text needs to reflect the kind of text a student could produce at this level. The errors need to focus on common errors at this level e.g. must to / must, said him / told him, married with / married to, bored / boring , the news is / are etc.

There are many other variations on the multiple choice theme, including choosing paragraph headings from a list of options, completing sentences by choosing from a list of phrases, choosing a picture to go with a taped description and labelling a diagram by choosing from a list of options, but all share their essential characteristics with the examples given above.

b) CANDIDATE-SUPPLIED RESPONSE ITEM TYPES

Items for which the candidate has to provide the response come in a variety of types from sentence completion or short-answer questions, where the candidate may have to supply as little as one word, a number or a short

phrase, to types requiring longer and more complex operations. It may be claimed that, by comparison with selection types, they:

- are easier to write
- allow for a wider sample of content
- minimize the effect of guessing
- allow for creativity in language use
- measure higher as well as lower order skills
- have a more positive effect on classroom practice
- can provide a similar degree of marking objectivity as selection items.

But these types of items also have their limitations. Even the most controlled types are difficult to construct in such a way that the required response is clearly indicated. There are often acceptable alternative responses rather than only one unambiguously correct response. This makes them time consuming and difficult to mark, often calling for examiner marking rather than clerical or computerized marking,

What these item types have in common is that, even in the most tightly controlled examples, candidates have to supply some language of their own, rather than choose among options supplied to them. Although they may appear in tests of reading, listening and structural competence, an **element of writing is involved**. In some cases this is done in the candidate's own language, but it is more likely to be a requirement of the test that the writing is in the target language. This raises the issue of **spelling**. The mark scheme must specify the policy towards spelling. If poor spelling is penalised, candidates for whom this is a particular problem will be unable to show their true level of language ability. If spelling is disregarded, there is the problem of deciding at which point a badly spelled version of a word ceases to be an acceptable representation of the word. Other concerns are **accuracy**, and the point at which a response which is not entirely correct may be considered acceptable or not. Problems of this kind make the process of assessment more subjective.

Short answer item

This item type consists of a question which can be answered in one word or a short phrase. The exact limits on the length of the answer should be specified. It is related to a text and generally used in tests of reading and listening.

Example:

The text, which is in a test of listening, is a man asking for information about a train.

TRAIN	
To:	Newcastle
Day of journey:	
Train leaves at:	
Return ticket costs:	
Food on train:	Drinks and ……
Address of Travel Agency	22 ……. Street

The correct response in this type of item is always one word or number, or a short phrase, not more than three or four words long. The range of the number of words expected should be made clear to the candidate. The questions must be written so that they cannot be answered from general knowledge or common sense alone.

In the above example spelling is tested explicitly in the last item. The street name (Mallet Street) is spelt out on tape (as it might be in real life) and the mark awarded only if the candidate spells the word correctly on the answer sheet.

ii) Sentence completion

In this kind of item part of a sentence is provided, and the candidate has to use information derived from a text to complete it. These items are used in tests of reading and listening.

Example:

The same text as in the example (man asking for information about a train) could be followed by completion items:

1. The man wants to travel on ……
2. His train leaves at …….
3. A return ticket costs ……
4. On the train the man can buy drinks and …….
5. The address of the Travel Agency is 22 …… Street.

As a general rule, blanks should be placed near the end of a statement, so that the candidate is provided with enough context to respond to the item.

When writing such items, it is important to ensure that there is either one unambiguous correct answer or a very limited number of acceptable answers which can be specified. The success of the marking process depends on this.

iii) Open gap-filling (cloze)

The varieties of cloze test have been described in the section on multiple choice and other selection item types. In an open cloze, the gaps are selected by the item writer, who focuses on the particular structures to be tested. The candidate's task is to supply the word which fills each gap in the text.

Example:

A NEW CRUISE SHIP

One (1) ….. the biggest passenger ships in history, the Island Princess, carries people on cruises around the Caribbean. More than double (2) ….. weight of the Titanic (the large passenger ship which sank in 1912), it was (3) ….. large to be built in (4) ….. piece. Instead, forty-eight sections (5) ….. total were made in different places. The ship was then put together (6) ….. these sections at a shipbuilding yard in Italy.

Open cloze works well in tests of **structural competence**. Prepositions and parts of verb forms can be deleted, for example, and there is often only one possible correct answer. Knowledge of vocabulary is more easily tested by means of a multiple choice cloze, as there are frequently too many possible correct answers to make an open cloze practicable. Gaps should occur approximately every seven to ten words.

iv) Transformation

In this type of item the candidate is given a sentence, followed by the opening words of another sentence which give the same information, but expressed through a different grammatical structure. For example, the first sentence may be active, while the second must be written in the passive. The candidate has to complete the sentence correctly.

This item type is used in tests of structural competence or writing at sentence level.

Example:

1. Maria missed the ferry because her car broke down. If..... .
2. Please do not smoke in this area of the restaurant. Customers are requested..... .

The crucial point to remember when writing this kind of item is that it will only work well if there is one clear correct answer (or a very limited selection of permissible variants). It is important to consider the number of testing points, and the acceptable answers for each.

For example, in the first item given above, there are three testing points, and one mark is available for each. The mark scheme is as follows:

	Marks
Maria's / her car had not broken down	1
Maria / she would not have missed /failed to catch the ferry	1

OR

Maria's / her car had not broken down	1
Maria / she would / could	1
have caught /got/ been able to catch the ferry	1

The mark scheme shows all possible acceptable responses. For the second part, there are many possibilities.

One way to tighten the focus of transformation items is to put the gap in the middle of the sentence so that the correct response is controlled by the structures on either side. Limiting the number of words that can be used (in this case to three) further limits the range of possible correct answers.

Example:

She moved to New York in order to find singing work.

She moved to New York ……wanted to find singing work.

because of / so / since she

v) Word formation

In this type of item one word is deleted from a sentence, and a related form of the word is given to the candidate as a prompt. For example, if the noun form 'singer' is required by the context of the sentence, the verbal form 'sing' is given as a prompt. The candidate has to supply the correct form of the word in its context.

It is used in tests of structural competence or writing where there is a focus on **testing knowledge of vocabulary.**

Example:

1. There were over fifty ..... in the orchestra. (MUSIC)
2. My parents …….. me to learn to drive. (COURAGE)

When writing items of this type it is important that the sentence gives an economical and unambiguous context to the target word. It tends to look more natural to put in a proper name rather than to use 'he' or 'she' all the time.

If the word formation items are set within a continuous text, more contextual support is given to the candidate than in the discrete items in the above examples. In the example below, the additional context allows the testing of a negative form in (2).

In Holland, people were so desperate to own tulip bulbs that their (1)	**BEHAVE**
….. became quite extraordinary. It was not (2) ……. for people to	**COMMON**
sell all their (3) …….. in order to buy a single tulip bulb. The situation	**POSSESS**
became so serious that laws were passed with the (4) ……. of	**INTEND**
controlling this trade in tulips.	

vi) Transformation cloze

As well as discrete items of the type shown above, the transformation item type can be used in a continuous text, creating a task which may be considered a variety of cloze procedure. It consists of a text with a word missing in each line, and a different grammatical form of the word required supplied. The candidate has both to find the location of the missing word and supply it in its correct form.

This kind of item can be used in tests of **structural competence or writing at sentence level.**

vii) Note expansion

In this item type the lexical components of each sentence are supplied in a reduced form which resembles notes. The candidate's task is to supply the correct grammatical form, including changes in word order and the addition of such elements as prepositions, articles and auxiliary verbs. A marker

indicates each point in the reduced sentence, where an addition or change should be made.

This type of task is most likely to appear in **a test of structural competence or writing.**

Example:

The following notes must be expanded to produce a letter written in reply to an invitation.

Dear Mr Harris

a) I be very pleased / meet you / teachers' conference / London last year.

b) It be kind / you / invite me / come and see you while I be / England / this summer.

c) I hope / pay a visit / your school / 26th and 27th June if / not be inconvenient.

d) Please / not rearrange / programme / me.

e) I be very happy / fit in / whatever you / do at that time.

f) I like / stay overnight / 26th June and hope / arrange accommodation / me.

g) I telephone you once / reach London / confirm / exact time / arrival / school.

h) I look forward / meet / again.

Yours sincerely

The comments made above on transformation items also apply here. The item writer must be clear about what the testing points are, and must construct a mark scheme as part of the item writing task. The above text contains forty specific testing points, each of which carries one mark. The mark scheme for the first sentence is as follows:

I was very pleased (1 mark) to meet you (1 mark) during/(while we were) at (1 mark) the (1 mark) teachers' conference (which was held) in London last year (1 mark).

A disadvantage of the item type is that it necessitates a rather complicated mark scheme, and is difficult to mark accurately.

viii) Error correction / proof reading

This task type consists of a text in which a word appears in an incorrect form in each numbered line. The candidate has first to identify the incorrect word, and then write it in its correct form at the end of the line. A simpler variation on this has the incorrect word already marked so that the candidate has only to supply the corrected form.

This is a tightly controlled type of candidate-supplied response item, most often seen in tests of structural competence.

In this example either the line is correct or there is an error, either of spelling or punctuation.

What colour can do for you	
Today, colour is a dazzling background to our lives in a way that our	1 ✓
ancesters can only have dreamed about. We take colour pictures of	2 ancestors
our holidays, watch colour TV go shopping in supermarkets which	3 TV, go
vibrate with colour and we have colour printers attached to our home	4
computers. We worry about the right colours for decorating the house	5
and we have favourites and pet hates where clothes are conserned.	6

In this task there is only one correct answer. It is important that the item writer knows the range of types of incorrect words to be used.

ix) Information transfer

Tasks described in this way always involve taking information given in a certain form and presenting it in a different form. For example, facts about a piece of equipment may be taken from a text and used in labelling a diagram

of that equipment, or figures given in a table or graph presented in the form of paragraphs of text.

This item type tests skills involved in **writing and structural competence.**

Example:

The input is an email written by a sales representative to his boss. The task consists of re-writing it as a report for Head Office. The 'transfer' referred to here concerns register from email (informal) to report (formal).

Email

John

Here's my impression of this year's sales conference. We were all very happy about where it was held – no problems getting to the centre. When we got there, the manager greeted us warmly and everything connected with the dinner that evening was pretty good. Unfortunately, we had problems the next day – it was incredibly hot and stuffy in the conference room so we felt very uncomfortable. Then they told us at lunch that there were lots of things on the menu we couldn't have but they didn't tell us why.

Report

The (1) …….. itself was found to be very satisfactory and (2) …….. to the centre was very straightforward. In addition, on (3) ……. we were given a very warm (4) …….. by the manager, and the dinner arrangements were perfectly acceptable. However, the (5) ……. in the conference room was extremely poor, causing great discomfort to all the delegates. As for lunch, many of the dishes were (6) …….. and no (7) ……. was given.

As with selection items, there are additional item types, including longer answers to questions and transfer of information involving maps, diagrams, graphs, etc., but the examples given should give an idea of the range of item types available which involve the candidate in supplying the language used in the response.

3) NON-ITEM-BASED TASK TYPES

i) WRITING: extended writing questions

Extended writing can be tested in a number of ways which vary in the degree of control exercised by the tester over the candidate's response. The amount of input can vary from tasks which begin with several hundred words of reading materials (e.g extracts from letters / articles / advertisements) to those which consist only of a title and indication of how many words to write.

In general, the greater the degree of control, the greater the amount of input and the heavier the reading load for the candidate.

Each of the ways of providing a stimulus for extended writing mentioned above has advantages and disadvantages.

Writing tasks with detailed input

Detailed input produces a more uniform set of responses from candidates, which makes marking quicker, easier and more reliable. It also means that reading as well as writing skills are being tested, which may detract from the objectives and raise questions about the validity of the task as a test of writing.

A more situationally authentic task can be set. Candidates can be provided with input in the form of letters or other documentation requiring a response directed towards a specific reader, with a register determined by the input.

Here is an example at B1 level.

- This is part of a letter you receive from your English penfriend.

  Help! It's my brother's 14th birthday next month and I can't think of a present to give him. What do teenage boys like getting as presents in your country?

- Now write a letter answering your penfriend's question.

Writing tasks with titles only

A stimulus which consists of an essay title only produces very varied responses, which are more difficult to mark fairly, but does not test comprehension of the input in any significant way. However, successful essay writing may depend on the candidate's background, culture and education, which may not be part of what the language tester is trying to assess.

One solution is to set a task which gives more input than an essay question, and includes the text-type to write, a reason for writing and sense of who the target reader will be. The response is more controlled than is possible if only a title is set, and markers can be given a mark scheme which outlines the likely content, together with criteria for assessing linguistic performance. Candidates are able to express personal ideas and to some extent to adapt the topic to their own interests.

This is an example at B2 level.

A British TV company is thinking of making a film about life in your area and has asked you to give them some information. Write **a report** describing the advantages of living in your area and saying how the area might change in the future.

The marker uses a mark scheme drawn up specifically for this writing task.

Content	Report should describe advantages of living in writer's area and possible changes in the future.
Range	Language of description, opinion & explanation. Vocabulary related to towns/countryside, housing etc
Organisation & Cohesion	Clearly organised with introduction & conclusion. Sub-headings an advantage.
Register & Format	Consistent register (neutral / formal). Formal report layout not essential.
Target Reader	Would be informed about writer's area.

A further problem is that of choice. If only one writing task is set, it is unlikely to be of equal interest or relevance to all candidates. If there is a choice of topics, this creates problems in comparing performance, especially if the responses involve a variety of text types.

One solution is to set two tasks, one compulsory and the other involving some choice. The compulsory task must be based on input which does not disadvantage any candidate, while the non-compulsory tasks can allow for special interests and the differing concerns of different age groups. The compulsory task will yield a more reliable score, while the aim of the non-compulsory task is to allow candidates to give their optimum performance, by offering them the opportunity to write on a topic of real concern or interest.

Examples of extended writing tasks at a variety of levels follow:

a) This example is taken from a test at an elementary level. There is a comparatively heavy reading input. The fact that the skills cannot easily be separated is recognized, as the test is described as a test of reading and writing. The expected response is uniform and easy to mark.

- Read this note from Jane.
- Then address the envelope to Jane's friend at the travel agency.
- The envelope is on the back of your answer sheet.

You asked me about the Happy Tours Travel Agency. It's in the centre of Belford. You'll see it at the end of Valley Road; the number is 104. My friend there is John Spratt. The post code is BL2 9XR and the telephone number is 0871 4266.

b) This task, although at the same level as the previous one, allows for greater freedom and exercise of imagination in the response.

You must see your friend, David, before tomorrow evening.
Write a note to David.

Say:

- Why you want to see him.
- Where and when to meet you.

Write 20-30 words.

c) This example is from a test at a higher level than the previous one, although still elementary. The combination of rubric and input produce quite a tightly controlled task, but the response is not totally predictable.

> You have just had a very difficult week with a lot of problems at home and at college. Using the information in the diary, write a letter of about 100 words to a friend, telling him/her about your bad week.
>
> | | |
> |---|---|
> | Monday | Washing machine broke down – water everywhere. |
> | Tuesday | No trains until 9 o'clock. Late for college. |
> | Wednesday | O.K.!! |
> | Thursday | Cat disappeared. |
> | Friday | English test - really difficult. |
> | Saturday | Broke one of grandmother's valuable plates - worth £500. |

d) The following example is from a test at an intermediate level, and represents a task with detailed input and a very tightly controlled response.

> Radford is a small industrial town without many facilities for its growing population.
>
> You are a member of the local council, which is meeting to decide how to use an empty site near the centre of the town.
>
> The four most popular ideas for the development are given below.
>
> **Using the information given**, complete the paragraphs on the next page, **giving reasons for your choices.**
>
> (Illustrations of a street market, a park, an arts centre and a sports hall follow)

Candidates complete three paragraphs. The opening phrase of each paragraph is provided:

a) I am most in favour of ……..……..

.....

b) My second choice would be .....

.....

c) However, I would not.....

.....

When devising this sort of very controlled writing task, it is important for the item writer to provide a model answer, and to specify which features of content or language use are to be rewarded, so that a workable mark scheme can be constructed.

e) The next task is from the same intermediate level test as example d), but represents a task of a contrasting type, a composition title which allows candidates free expression of their own opinions and imagination. This pair of tasks highlights the range of demand made on assessors of writing tasks.

You recently took part in a class discussion about choosing an interesting job. Your teacher has now asked you to write a composition, answering the following question and giving reasons for your choice.

Would you rather be a politician, a teacher or a musician?

f) The next example consists of a task which requires candidates to process about 400 words of input material, and use the information appropriately to perform the task required. Candidates must read all the input material carefully, selecting that which is important. Input material may consist of varied combinations of texts and notes, sometimes supported by illustrations or diagrams. The task is often divided into more than one section. Task types will vary in Part 1, and may include formal letters, informal letters, reports, articles, notes or any combination of these.

Part 1 (CAE)

1 You are a member of the student welfare committee at your college. The committee has recently received a memo from the college Principal, announcing major cuts to the existing sports facilities. The committee is opposed to these reductions and has prepared the poster shown below.

You have offered to write an article for inclusion in the next college newsletter in order to ensure that students have as much information as possible, and to get their support. You have also been asked to reply briefly to the Principal's memo, informing him of the results of the survey and making him aware of the committee's opposition to the cuts.

Read the poster below and the Principal's memo on the next page, to which you have added your comments. Then, **using the information carefully**, write the article and letter as instructed on the next page.

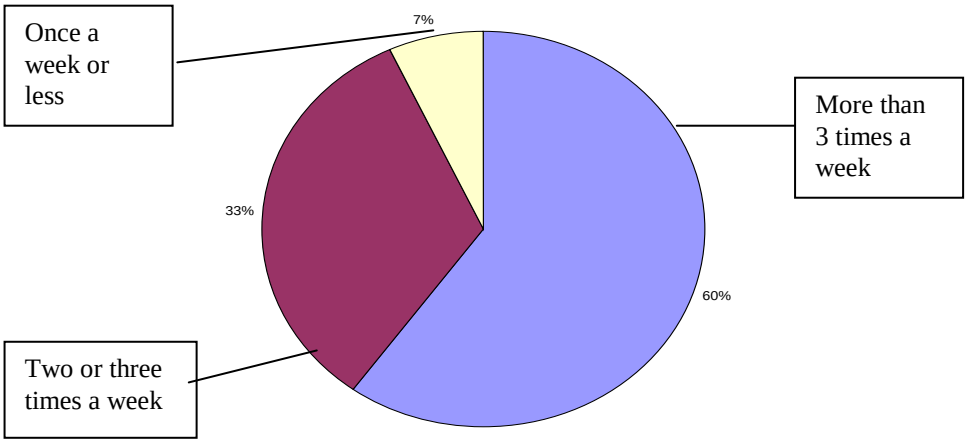
S – O – S
Save our Sports!

This college is planning to

- Sell off the sports field
- Hire out the sports hall to the public

How do you feel about this?

Our survey shows that most students use these facilities frequently:



We must fight this decision together!

To: The Student Welfare Committee **From: E G Baton, Principal**

REVIEW OF SPORTS FACILITIES

We didn't!

As you know, the college has been forced to consider various cost-cutting measures recently.

In connection with this, my colleagues and I have been discussing the college sports facilities. In our view, these are very much under-used. We have estimated that a small number of students use the sports hall, and the sports field seems to be used only twice a month for football matches.

Not true – NB our survey

What about practice?!

Due to this apparent lack of student interest, I have decided on some immediate changes. I intend to make the sports hall available for hire to the public during the day. I would also like the committee to introduce a student membership charge – students will be required to pay for these facilities.

All day?

Unfair

Finally, the sports field has been put up for sale. At least one buyer is interested in the site for building development.

Never!!

I trust the committee will support my decision.

E. G. Baton

Principal.

Now write:

a) an **article** for the college newsletter (approximately 150 words)

b) an appropriate **letter** to the Principal (approximately 100 words). You do not need to include postal addresses.

You should use your own words as far as possible.

g) The final example is of a task from the same test as example f). This task, while providing more detailed input than that given in an essay title, allows for some freedom of response, while specifying clearly the constraints within which candidates have to perform. The fact that the topic might not be equally attractive or relevant to all candidates is not a problem, as this task is one among a choice of four.

You are a member of the students' committee at a college where there are many students from all over the world. The college has a lot of sports facilities which students can use in their spare time. There is also a wide range of recreational activities which students can participate in. The committee is concerned that many students are not taking advantage of all there is on offer.

You have been asked to write a leaflet which:

- informs the students about the facilities and activities that are available
- points out the benefits of taking up these opportunities
- encourages students to use the facilities and join in the activities

Write the text for the leaflet.

ii) Speaking tasks

The issues connected with testing speaking skills tend to focus on the training, roles and reliability of examiners rather than on the speaking task. However, there are a number of features of procedures and tasks (rather than examiners) which have to be considered when a test of speaking is planned.

These features include:

- whether the test is administered face to face or in a language laboratory

- whether role play is used, or candidates are asked only to speak about themselves and their own views and opinions

- whether prompts are verbal, written, pictorial or some combination of these;

- how many candidates are assessed at any one time, and whether they are asked to interact with each other, or only with the examiner.

As with the testing of extended writing, the degree of control imposed on the response by the input varies, and in general the more complex input produces the more controlled response, which is easier to assess.

At its simplest and least controlled, the test of speaking is a series of questions from the examiner to the candidate, which follows a prescribed list of topics. It probably begins with a 'warm up' consisting of a few general personal questions from the examiner. This kind of interview cannot really be said to consist of tasks.

The more tightly controlled type of speaking test is one in which the examiner sets the task by providing a rubric (in the form of oral instructions) and input (in the form of pictures, text, etc.) and then giving the candidate(s) a certain amount of time to produce a response. The language of examiners, as well as that of the candidate, may be subjected to considerable control, as they may be asked to use a form of script known as an interlocutor frame, which ensures that the interview takes the same basic form for all candidates, even though individual tasks may differ. This is discussed further in Module 4, page 171.

The choice of task type is to some extent determined by whether the candidate is being tested individually, in a pair, or in a larger group. In pairs or groups collaborative discussion and problem solving tasks are possible. The series of tasks presented to candidates may or may not be linked by a common theme. Some examples of the range of task types used in face to face controlled interviews are given below.

a) Presentation

Candidates are asked to speak for a specified number of minutes on a topic which has either been prepared in advance or is given shortly before the test. This technique allows candidates to talk on topics of real concern and interest to themselves. The disadvantage is that speeches on mainstream topics may be prepared and / or memorized.

If more than one candidate is being assessed, questioning each other on the content of their presentations and answering questions may be part of their task. This ensures that each candidate listens actively to the other and tests their interactive communicative skills.

The following is an example of the instructions given to candidates for a presentation task.

In this part of the test you are each going to talk on your own for about two minutes. You need to listen while your partner is speaking because you'll be asked to comment afterwards.

I'm going to give you a card with a question written on it and I'd like you to tell us what you think. There are also some ideas on the card for you to use if you like.

A
How are young people's lives changing?

- learning methods
- opportunities
- expectations

B
What role can old people play in society today?

- the family
- the workplace
- public life

In a test of business language at advanced level, the candidate would need a wide range of business vocabulary and knowledge to deal with more specialist presentation topics:

A

Customer Relations: the importance of making customers feel valued.

B

Strategic Planning: how to decide whether to purchase or rent company premises.

C

Finance: how to decide whether to float a company on the stock market.

b) Use of picture prompts

Candidates look at a photograph or set of photographs. Depending on the level, they may be asked to describe what they can see, or what is happening, to suggest a common theme for the photos or to give opinions on the situation depicted. A discussion may be developed out of the theme introduced in the photos.

Example:

In Part 2 of the June 2004 FCE exam candidates are shown two photos: one of a mountaineer camping on the top of a mountain and one of people whitewater rafting.

The following is an example of how these pictures would be used:

Interlocutor: Now, *(Candidate A)*, here are your two photographs. **They show people enjoying the natural world**. Please let *(Candidate B)* have a look at them.

Hand over picture sheet to (Candidate A).

I'd like you to compare and contrast these photographs, and **say why you think the people have chosen to go to these places.**

Remember, *(Candidate A)*, you have only about a minute for this, so don't worry if I interrupt you. All right?

1 minute

Candidate A:……

Interlocutor: Thank you. *Retrieve photographs.*

(Candidate B), **would you like to spend time in either of these places?**

Approx. 20 seconds

Candidate B……

Interlocutor: Thank you.

c) Written prompts

Discussions can be prompted by a brief written statement, for example:

> Today's news events are tomorrow's history

Candidates are invited to react to the statement and suggest some events of national or international importance that have happened during recent times or during the candidates' lifetime which they think have affected or will affect the course of history. With minimal input such as this, the task is very demanding.

If candidates are being examined in pairs, then collaborative tasks can be set. Candidates can be given oral instructions and visual stimuli to form the basis of a task which they do together. This tests how well they can interact with each other (listening / turn-taking / initiating / responding).

Here is an example from a test at B2 level.

> I'd like you to imagine that a busy international hotel is looking for staff for the holiday season. Here are some of the jobs available.
>
> *Candidates are given picture representing 7 different jobs.*
>
> Talk to each other about how difficult it would be to do these jobs without training. Then say which job you would find the most difficult.

d) Information gap tasks

A real need for communication is often built into speaking tasks, particularly where two candidates are being assessed together. This often takes the form of a discovery or problem solving activity of some sort, often prompted by pictures. Some variations on this are:

- Each candidate has a picture the other cannot see. Although similar, they are not identical, and one has to describe the pictures to the other to discover what the differences are.

- Each candidate has a set of pictures which are identical but arranged in a different order. One candidate must describe one of the pictures, and the other must identify by number which picture was being described.

Such tasks work best at elementary level as the language produced will be simple and to some extent repetitive.

4) RUBRICS

The definition of **rubric** given in the Introduction p.7 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.

The rubric must present as clearly as possible the task which the examiner is setting the candidate. There should be no room for confusion, no need for clarification, as both are likely to create anxiety in the candidate and impair his/her performance and affect the reliability and validity of the test.

In **listening tests**, the rubric should be printed on the question paper and also recorded onto the test recording.

In **speaking tests** (face-to-face interviews) the rubric will be oral instructions from the interviewer/interlocutor/examiner. There may be an opportunity for a candidate to ask for clarification of the task and this is justifiably part of the interaction.

Here, as an example, is the rubric from a writing task in the First Certificate in English:

> A British TV company is thinking of making a film about life in your area and has asked you to give them some information. Write a report describing the advantages of living in your area **and** saying how the area might change in the future.
>
> Write your **report**.

The key questions to ask about a rubric are as follows.

How clear is it? (Is it possible to misinterpret the nature of the task?)
How easy to understand is it? (Is the language used at an appropriate level? This is particularly important in language testing at the lower levels.)
How adequate is it? (Is *all* the necessary information given?)
How relevant is it? (Is *only* necessary information given?)
How consistent are rubrics? Throughout the test the language of rubrics should be standardized, 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 writing of rubrics is an important part of the overall writing process. Item writers should use a checklist such as the following:

- 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? In a test of reading, this means that the level of language used in the rubric must be *below* the level of language being tested.
- Is the language simple and clear?
- Is there any superfluous language?
- Are there any double negatives?
- Is there any room for ambiguity or misunderstanding? It is useful for item writers to get a colleague to try out the item, as writers are often too 'close' to the material to spot all possible problems.
- Does the rubric contain all the necessary information and limitations?

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
- clear indications 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

This detailed information should be clearly defined in Guidelines to item writers for future use and consistency. The item writer can set up templates to facilitate the writing of items.

5) KEYS, MARK SCHEMES AND RATING SCALES

Any test task must be accompanied by:

- an appropriate **rubric**
- an accurate **key (**set of correct answers for objective item types)
- a **mark scheme** with all acceptable answers for productive tasks (short answer items or writing tasks)
- a **rating scale** for subjective tests with set of task requirements and marking criteria.

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?
- Are there a sufficiently limited number of possible answers? If there is too little restriction on possible answers then marking may prove problematic. The test may well be marked by someone who is not a language specialist.
- 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.)

Various methods of arriving at a score for a speaking or extended writing task are described in Module 4. Generally this involves use of a rating scale, which may, for example, break down the skills being assessed in a test of speaking into areas such as pronunciation, intonation and accurate use of structure, and mark each on a scale of 1-5. To help the assessor, a brief description is provided of a typical performance at each level. The final score for a speaking task or for a writing task may be arrived at by giving a mark out of 5 on a set of, for example, six scales, so that a total score for the task is a mark out of 30.

A slightly different method of assessment can be related to the examples of writing tasks f) and g) given above. As they are both tasks which are, to a differing extent, controlled by input, a **task-specific mark scheme** can be constructed for each. Under the headings of **content**, **organisation and cohesion**, **range**, **register**, the impression made on the **target reader** and **accuracy of language,** short descriptions are given of the standard of performance which would achieve the score of 5 for the whole task. A separate **general mark scheme** gives a more generalized description of optimum performance at each level. Task-specific mark schemes are produced by the item writer and confirmed during trialling, based on the reading and marking of a range of candidates' scripts.

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. 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.

EXERCISES

Materials for the Guidance of Test Item Writers

1. There is an identifiable fault in each of the following items.

Describe what is wrong with each item, and then decide whether you can re-write it in such a way as to make it an acceptable item, or whether you would reject it.

a) *In the story, the merchant was unhappy because it*

A *rained.*

B *was windy.*

C *was dark*

.

D *was windy and rainy and he had forgotten his overcoat.*

b) *A tea* *B tee* *C party* *D dinner*

c) *A ran* *B fast* *C runs* *D is running*

d) *What is the physical state of sodium at room temperature?*

A *fluid*

B *gas*

C *liquid*

D *solid*

e) *What happened in the period 1980 - 1985?*

A *The birth rate rose.*

B *The birth rate fell.*

C *The birth rate remained unchanged.*

D *The death rate fell.*

f) *They said they*

A *had gone.*

B have went.

C had go.

D had went.

g) *When is it not appropriate not to be absent from class?*

A When you are sick.

B When you are young.

C While class is in session.

D Whenever the teacher is angry.

h) *We learn from this passage that Napoleon was*

A British.

B French.

C Polish.

D German.

i) *The boy took the newspaper*

A because he wanted to read it.

B because he wanted to wrap a gift in it.

C because he wanted to dispose of it

.

D because he wanted to remove an article.

j) *At the end of the lesson the teacher..... up the children's books.*

A picked

B collected

C took

D pulled

k) *Martha arrived late for her interview because*

A *she overslept that morning.*

B *her train arrived late.*

C *of the train drivers' strike.*

D *she missed her train.*

l) *Why did Mr Young go to the bank?*

A *to get a new cheque book*

B *to buy a sandwich*

C *to ask for some advice*

D *to pay in a cheque*

m) Which magazine should you buy if your hobby is gardening?

A *Beautiful Homes*

B *At Home With The Stars*

C *Home and Family*

D *Homes and Gardens*

2.

Read the four test items and then do the task which follows them.

1.

My village is called Combton. There are only three shops in my village. There is a bakery, a butcher's and a newsagent's. The bakery is open every day from Monday to Saturday. It opens at eight o'clock every morning and closes at four o'clock. The butcher's opens the same hours on the same days, but it is closed all day Wednesday. The newsagent's is open from seven o'clock every morning Monday to Saturday. It closes nearly every day at five o'clock. On Wednesdays it closes at one o'clock.

Which shops are open in Combton on a Saturday morning?

Answer: All three; the bakery, butcher's and newsagent's.

2.

By now I've got the whole place clearly charted.
Our garden, first: where I did not invent
Blinding theologies of flowers and fruits,
And wasn't spoken to by an old hat.
And here we have that splendid family

I never ran to when I got depressed,
The boys all biceps and the girls all chest,
Their comic Ford, their farm where I could be
'Really myself'. I'll show you come, to that,
The bracken where I never trembling sat,

What, in your own words, is Philip Larkin, the poet writing about?

3.

Friends, Romans, countrymen, lend me your ears;

I come to bury Caesar, not to praise him.

The evil that men do lives after them,

The good is oft interred with their bones;

Who spoke these words?

4.

A man is returning home from a day at the market. He has bought a chicken, a fox and a bag of grain. Don't ask me why he bought a fox! He comes to a river. It is too deep to wade across and too rough to swim. There is a small boat moored at the bank, but it is too small for the man and all his possessions. He can only take one thing across at a time. He can row backwards and forwards as many times as he likes, but if he leaves the fox and the chicken alone, the fox will eat the chicken. If he leaves the chicken and grain, then the chicken will eat the grain. How can he get everything across the river?

Answer:

i) He takes the chicken across and leaves it.

ii) He comes back and takes the fox across and leaves it.

iii) He brings the chicken back with him.

iv) He takes the grain across and leaves it with the fox.

v) He comes back for the chicken.

Look back over the texts and questions and think about the purpose of each item, and the difficulty of each item.

Can you add anything to this list of points illustrated?

- linguistic complexity of the text or stimulus
- linguistic complexity of the response required
- cognitive complexity of the task
- world knowledge required not language ability
- linguistic conventions of the text type

Module 4

MODULE 4

ISSUES IN MARKING AND SCORING

1. PROVIDING A FAIR RESULT

Item writers may well feel, once items have been edited and pretested and banked for use in live tests, that their role in the testing process has come to an end. What happens to a test once it has been administered, how it is marked and the raw score converted into a grade which can be reported to the candidate, is a wide topic and one which may not seem very relevant to the interests of the item writer.

How is the item writer involved in the issues related to marking and scoring a test?

The item writer's role is to provide well-designed items and unambiguous mark schemes. This has a significant impact on the ease with which an item can be marked accurately and mistakes in marking (and the unfair results which follow from this) avoided. Directly or indirectly, the item writer has an impact on the *entire* testing process.

The item writer needs to understand the subsequent stages of the testing process so that feedback from live tests on how items perform, and mark schemes work, is meaningful and can inform the writing of future items. This also contributes to the development of the test itself, as broader issues will emerge about how a particular skill is being tested.

It is important to have such **a feedback system** in place. This might be done regularly at editing meetings, for example, or by written feedback to the item writing team. Team feedback may be more appropriate than individual feedback since the items have been through the editing process with input from more than one item writer.

What are the main issues in marking and scoring?

In any kind of test, the test provider's essential task is to give **a fair result** to the candidate. The aim is always to achieve fairness by an insistence on accuracy and consistency and by keeping scoring error to a minimum.

One aspect of this emphasis on fairness is **reliability**. This is an issue of particular importance in language testing and this module summarises a number of issues related to reliability.

In terms of methods of marking and the particular problems or issues to be considered, the main classification is the following:

- tasks with a format which allows some form of **objective** marking
- tasks whose nature demands a more **subjective** judgement

This module also considers further issues related to the marking of objective and subjective tests.

2. RELIABILITY

a) In objective tests

Reliability in testing is often defined as consistency of measurement i.e. an assessment procedure should measure consistently that which it sets out to measure. A candidate taking two versions of the same test on two occasions close in time to one another would be expected to get approximately the same score, and not find that one version was noticeably more difficult than the other.

What are the factors affecting reliability in objective tests?

i) Internal factors

These concern features such as the number and quality of the items used. There are various statistical methods for checking test reliability in terms of internal consistency. These generate statistics in the form of coefficients; the most commonly used are Cronbach's alpha (coefficient alpha) and Kuder-Richardson 20 (K-R20). The formulas for these coefficients can be found in Appendix A. Calculations of this kind are done by computer using statistics software such as SPSS or item analysis software such as MicroCAT.

ii) External factors

External factors which can affect reliability concern:

- external conditions

The score achieved could be affected by the quality of the room in which the test is taken (lighting, comfort, space, noise levels from outside). The person administering the test may give clear and explicit instructions or may be unsympathetic and unsupportive. Such factors are avoidable and should be dealt with in training and preparation procedures for the administration of the test.

- attitude and behaviour of candidates

The score achieved could be affected by the extent to which candidates feel confident or nervous, healthy or off-colour as well as by candidates who are uncooperative and unsympathetic to the idea of being tested – a reluctance to expand on answers in a speaking test, for example. The reliability of the result will also be affected by candidates who guess the answers in multiple choice tests and by candidates who have had intensive practice or no practice at all with the format and item types used in the test.

b) In subjective tests

The question of reliability in subjective tests concerns the quality of the judgements made by markers and examiners. What is aimed at is:

- **inter-rater reliability**
  consistency of judgement between different markers of the same test
- **intra-rater reliability**
  consistency of judgement by the same person on different occasions and under differing circumstances

Clearly this type of marking is far more influenced by human changeability and error than computerised or clerical marking. It is, therefore, unrealistic to expect total reliability. There are, however, many measures, the rigorous training of examiners, for example, which can be put in place to make subjective tests as objective and reliable as possible.

These measures are dealt with below in the section on marking subjective tests (writing and speaking) on pages 162-176.

3. SOME ISSUES IN MARKING AND SCORING OBJECTIVE TESTS

Marking the candidate's response

Reading, listening and structural competence are commonly tested by means of item-based tests. The item types used are commonly selection types such as multiple choice, true/false, gap-filling and matching and the more tightly controlled of the candidate-supplied response item types, such as short answer items.

For these types of items, there is often only one possible correct answer (the key) as in multiple-choice or true/false items, or a very limited and easily defined number of acceptable answers (which form the mark scheme) such as in grammatical transformation or cloze tests.

How can objective tests be marked?

i) The marking of **tests with keys** can be done by clerical markers or by computer as there is no requirement for any discretion in marking to be exercised.
ii) The marking of **tests with mark schemes** can be clerically marked (as opposed to examiner marked).

Computerised Marking

The marking of objective tests with keys can be marked by computer. The use of technology usually involves some kind of optical scanning. This has been available for many years in the form of **optical mark readers** (OMRs) which are used with specially designed OMR answer sheets (a specimen OMR sheet is included as Figure 1). More recent developments have seen the use of bar code labels in the context of examining.

In the case of the items with only one correct answer (such as multiple choice), the information can be scanned straight from the candidate's mark sheet by an optical mark reader, and scores totalled by a computer. Software is also increasingly being used to mark single words or phrases with a narrow range of acceptable answers (information transfer – form filling, for example) where the candidate is instructed to write each letter (usually in capitals) or numeral in a separate box on the OMR answer sheet.

The use of computerised methods of marking means that the final score given for this kind of test can be arrived at with a high degree of accuracy.

Clerical Marking

The marking of objective tests with mark schemes can be done by clerical markers. This means that the markers – unlike examiners - do not need any high degree of specialised skill or experience of language tests to mark these tests. The marker has to follow a mark scheme which sets out all the acceptable answers and the number of marks to be given for each item.

What is **the role of the item writer** in this process?

It should be part of the item writer's job to look ahead to how the test will be marked, and avoid submitting items which would be sure to cause marking problems.

In an open cloze test, for example, it is undesirable to create gaps in a text which can be filled by a large number of words, each of which could be correct.

A wonderful new (1) …….. has just been published.

In the example above, the number of nouns which could fit the gap is considerable (book, magazine, edition, title, song, picture, report, work, CD-rom etc). While potentially useful as a vocabulary-building exercise in class, it is unsuitable as a test item (and would clearly be rejected at editing).

It is also vital that the item writer provides a clear and accurate mark scheme when the items are first written, ensuring that the range of acceptable answers is limited. All acceptable answers should be available to the clerical marker on the mark scheme, so that the marker is not asked to make decisions about what is or is not acceptable.

In the example below, three transformation items are given, followed by the mark scheme.

(a) Please do not smoke in this area of the restaurant. Customers are requested .....
(b) Although he took a taxi, Bill still arrived late for the concert. In spite .....
(c) Carol finds it easy to make friends. Carol has no .....

MARK SCHEME
(a) not to smoke OR to refrain from smoking
(b) of taking/having taken a taxi the fact that he/Bill hired/had hired/took/ had taken a taxi he/Bill (still) arrived late.
(c) problem(s)/difficulty/trouble (in) making friends/forming relationships

It is important for candidates to know what is expected of them. In items (b) and (c) above, the range of possible answers is such that the candidate could be uncertain – and waste time wondering about which of the correct answers is *really* correct.

One way to tighten the focus of a transformation item and limit the range of possible answers is to give the beginning and end of the sentence and limit the words in the gap. In the example below, the gap must be filled with between two and five words, one of which must be the key word given.

(a) Mike's father started the company that Mike now runs. **set**

The company that Mike now runs …… his father.
(b) I don't recommend hiring skis at this shop. **advisable** It's …….. skis at this shop.
(c) 'This is the best hotel I've ever stayed in,' my colleague said. **never** 'I've …… hotel than this,' my colleague said.

MARK SCHEME
(a) was **set** up by
(b) not **advisable** / advisable not to hire (your)
(c) **never** stayed in a better

Further examples: a complete set of items and accompanying mark schemes taken from the Use of English component of the First Certificate in English can be found in Appendix B.

Issues related to clerical marking

i) Expanding the mark scheme

Item writers cannot always predict all the responses that candidates will produce. It is common when marking short answer items for markers to find

answers given by candidates which appear to be acceptable but which were not anticipated. The mark scheme then has to be adjusted.

In a school internal test it is relatively easy to deal with this situation. In a national / international examination clerical markers are often organised into teams with a Team Leader or Co-ordinating Examiner with specialist knowledge of the test to refer to when decisions over what is acceptable have to be made. The mark scheme may need to change several times during the process of marking, to include further acceptable answers.

ii) Training and monitoring

The training of clerical markers needs to focus on the correct application of the mark scheme and the need for accuracy in marking. Errors may still occur, particularly when large numbers of candidates are taking an examination. Markers may misread mark schemes or the candidate's handwriting, write down the wrong number of marks or add up the marks for a section or paper wrongly. This needs to be addressed in training. Markers also need to be monitored, as with Team Leaders making ransom checks on samples of each marker's work.

Turning raw scores into examination results

The score for an item-based test is an accumulation of correct responses. This raw score has next to be transformed into a result which can be interpreted meaningfully by the candidate and by any other individual or institution, such as an employer or educational establishment, wishing to use that result as a measure of ability. A score of, for example, 27 marks out of 50 cannot be interpreted until it has been related to a meaningful reporting system. Some familiar examples of reporting systems are pass/fail, or Distinction, Merit, Pass or A,B,C.

4. SOME ISSUES IN MARKING AND SCORING SUBJECTIVE TESTS

Marking the candidate's response

Performance tests i.e. tests of productive skills (speaking and writing) tend not to be item based, although writing can be to some extent tested by means of candidate-supplied response items. These tests make different demands on markers, and the process of examining these skills is generally regarded as more subjective.

How can the assessment of subjective tests be fair?

Factors which affect the quality of a subjective test relate to:

i) Materials and test format

One way to make subjective tests fairer to the candidate is to ensure that the test is composed of **a variety of tasks**. The ability level of the same candidate may vary widely according to the topic or kind of task set so any test of writing or speaking should be composed of more than one task. An element of choice in the test also gives people of different ages, backgrounds and interests some chance of finding a task which allows them to perform at their best. A candidate may prefer to write narrative than the factual text required for a report, or to discuss the effect if climate on people's behaviour to the effects of tourism, depending on the candidate's country of origin.

NB Giving a choice of tasks can make it difficult to compare performances on different tasks, especially if different text types are being produced. One approach is to set one compulsory task followed by a choice of task(s).

ii) Assessment procedures

Ensuring fair assessment in performance tests will depend on the following aspects of the process:

- **recruitment and induction of markers and examiners**

Skilled markers and examiners are needed for performance testing and, for widely taken national / international tests, these are often teachers familiar with the examination through preparing students to take the test. Training sessions give them a chance to become thoroughly familiar with the scale or list of criteria against which candidate performance is to be assessed.

- **sound standardisation procedures**

The training of markers and examiners includes a process referred to as 'standardisation', the aim being to enable them to provide a standard level of judgement of candidates' output.

- **clear marking criteria**

The scale may consist of numbers, letters or 'labels' ('Good', 'Adequate' etc) and contain statements of what each point on the scale refers to. These are the scale descriptors. Markers and examiners need a thorough understanding of the principles behind the particular scale(s) they are working with.

- **monitoring and evaluation of markers and examiners**

The performance of markers and examiners must also be assessed. It is necessary to set up a system of checking, monitoring and evaluating their performance and provide constructive feedback.

- **number of markers involved in each assessment**

The question of how many markers or examiners take part in each assessment affects tests of both writing and speaking.

This module will look at these issues for tests of writing and speaking.

a) TESTS OF WRITING

When a test is being designed, decisions need to be made about whether the writing test is to test primarily mechanical accuracy (e.g. accuracy of grammar, punctuation and spelling) or whether the achievement of some kind of task which reflects an authentic use of written language is the main criterion (e.g. writing a letter of complaint about a problematic holiday). Depending on the decisions made at this stage, the input for test questions may be simply composition titles or several pieces of realia such as advertisements, letters and excerpts from newspaper articles.

The main issue with assessment of tests of writing is how best to control the subjective element in order to achieve fair assessment. Depending on the sort of input provided and the intentions which underlie the test, scripts may be marked for **mechanical accuracy**, **analytically marked** according to detailed criteria or **holistically marked** (impression marking).

Methods of marking

i) Marking for mechanical accuracy

The marker may underline every error of grammar, spelling and punctuation and assess the piece of writing according to how many errors it contains. One way of doing this is to fix a maximum score and deduct marks from it according to how many errors are made. Another way is to allocate a proportion of the marks for grammatical accuracy, spelling etc and deduct marks for the type of errors which occur.

The issue with this type of marking is that it penalises the more ambitious candidate who writes more and uses more complex language, and rewards the candidate who produces a briefer, simpler text. It is difficult to adapt this system so that it is able to take account of the register of a text or the overall communicative impact of what has been written. Grammatical accuracy and spelling can be tested in objective item-based tests.

ii) Analytical marking

The marker may be asked to give a mark on a scale (for example, 1 to 5, where 1 is a poor performance and 5 excellent) after assessing separately various aspects of the writing, such as range of vocabulary, use of structure and appropriate use of register. The final score is a composite of all the subscores or a profile. Normally, various aspects of the range and accuracy of the language are assessed, together with an assessment of how well the set task has been achieved.

iii) Holistic or impression marking

The marker gives a mark based on an assessment of the overall effectiveness of the piece of writing. With this kind of marking there is no breakdown into separate marks for separate aspects of writing skill, but some criteria still have to be kept in mind.

The marker may give a mark on a scale of 1-5 or 1-20, for example, based on several very quick readings of the script, but backed up by an internalised concept of the level of task achievement, use of grammar, structure and vocabulary required for the language level of the test. A second independent marking of each script may be desirable. Impression marking can also be done by indicating the errors on the script and ticking examples of very good language to enable the marker to establish the 'best fit' to the descriptors on a scale of 1-5.

These are two examples of mark schemes for assigning an impression mark for free writing tasks in one test of English at an intermediate level (B2).

A

18-20	Task successfully carried out with a wide range of expressions and minimal, if any, errors.
16-17	An ability to produce more than a collection of simple sentences with only occasional lapses; task successfully carried out.
11-15	Simple but correct realisation of task with some errors which do not distract from the content; appropriate credit for range.
8-10	Message communicated, but errors noticeable; attempt at task not entirely successful.
5-7	Lack of language control shown by frequent basic errors; task only partly realised/rubric neglected.
0-4	Language breakdown; content irrelevant or too little for assessment.

In mark scheme A there is no attempt to define what a mark of, for example, 12/20 or 6/20 means. On a 20-point mark scheme it is difficult to make such fine distinctions. The consistency with which the scale is used, therefore, will depend heavily on the skill and training of the markers.

B

5	Full realisation of task set. • All content points included with appropriate expansion. • Wide range of structure and vocabulary within task set. • Minimal errors, perhaps due to ambition; well-developed control of language. • Ideas effectively organised, with a variety of linking devices. • Register and format consistently appropriate to purpose and audience. Fully achieves the desired effect on the target reader.
4	Good realisation of task set. • All major content points included; possibly one or two minor omissions. • Good range of structure and vocabulary within task set. • Generally accurate, errors occur mainly when attempting more complex language. • Ideas clearly organised, with suitable linking devices. • Register and format on the whole appropriate to purpose and audience. Achieves the desired effect on the target reader.
3	Reasonable achievement of task set. • All major content points included; some minor omissions. • Adequate range of structure and vocabulary, which fulfils the requirement of the task. • A number of errors may be present, but they do not impede communication. • Ideas adequately organised, with simple linking devices. • Reasonable, if not always successful attempt at register and format appropriate to purpose and audience. Achieves, on the whole, the desired effect on the target reader.
2	Task set attempted but not adequately achieved. • Some major content pointed inadequately covered or omitted, and/or some irrelevant material. • Limited range of structure and vocabulary. • A number of errors, which distract the reader and may obscure communication at times. • Ideas inadequately organised; linking devices rarely used. • Unsuccessful/inconsistent attempts at appropriate register and format. Message not clearly communicated to the target reader.
1	Poor attempt at the task set. • Notable content omissions and/or considerable irrelevance, possibly due to misinterpretation of task set. • Narrow range of vocabulary and structure. • Frequent errors which obscure communication; little evidence of language control. • Lack of organisation, or linking devices. • Little or no awareness of appropriate register and format. Very negative effect on the target reader.

In mark scheme B there is more explicit definition of each criterion, allowing the marker both to assess various aspects of the language and the overall communicative effect of what has been written. In this way it is possible to give credit to candidates who, despite producing errors, write using an

ambitious range for the level or produce a text-type in a culturally unique and/or creative way.

Task Fulfilment

Examiners involved in impression marking may be guided by a general impression mark scheme, (see mark schemes A and B above) which is not related to any specific examination question, but provides detailed descriptors of the level of performance expected for each mark which could be given. They may also be given a mark scheme specific to each individual task with a detailed description of the elements of task achievement, register, range of structure and vocabulary expected, etc. This is to ensure a consistent approach to assessing how well the task has been achieved.

Examiner training

In general, the less analytic and mechanical the method of marking, the more highly skilled and trained examiners need to be. It is important that they have some knowledge of the examination level and content, probably from their experience of preparing candidates for it.

What are the stages in training examiners?

For an internal school test the procedures below may be modified but the essential steps in the process are the same, regardless of the numbers of candidates.

i) Standard setting

Once the test has been taken, the chief examiner looks at as many scripts as practical in the time available to gain an overall impression of what has been written and whether the tasks set have been successful or caused any problems. After selecting a number of scripts representative of each level on the rating scale being used, the chief examiner meets with senior examiners to mark the scripts and discuss and compare judgements. Issues related to how task fulfilment is to be approached are also discussed. Once marks have been agreed, the reasons are noted. These scripts can then be used for training examiners.

ii) Standardisation

Examiners are given the standardisation scripts to mark and then discuss their scores with a senior examiner or team leader (either at a meeting or by telephone). If they are marking too high or too low, they can adjust to the level of the examination. It is useful if the reasons for the marks are available in written form for examiners to refer to during the marking period, if necessary. It is important that examiners go through this standardisation process each time they mark.

Appendix C consists of a writing sub-test followed by some sample scripts and the kinds of comments used in assessing them and guiding the assessments made by examiners.

NB Examiners marking for the first time will benefit from a separate induction meeting to introduce them to the 'mechanics' of marking (how to indicate errors, pace of marking, timescales for returning scripts, support available from senior examiners etc).

iii) Monitoring examiners

During the marking period an examiner's marking can be monitored for accuracy of level and task assessment by a senior examiner or team leader. It is helpful in the early stages to 'steer' the examiner's assessment of the language level by reference to the mark scheme. Extended intervention is not desirable since examiners may react by changing their marks too frequently and not establishing their own consistency.

iv) Evaluation of examiners

When the marking is completed it may be possible to produce statistical evidence of an examiner's consistency in marking. It is useful to provide this feedback to the examiner for future reference.

At the end of the marking period the chief examiner writes a report on the writing test with information about how successful the tasks set were. It is important that this feedback goes to the item writers of the test to inform the quality of items written in the future.

Multiple marking

Multiple marking improves reliability. Writing tasks which are impression marked are also likely to be double marked, so that every piece of writing is assessed by more than one person. If two people give widely varying marks, the script is marked a third time, possibly by someone like a team leader or the chief examiner.

This kind of marking may be done as a 'table-top' session, with all the examiners gathered together to mark all the scripts from one administration of the examination. This makes it possible to monitor the speed and accuracy of marking as it takes place, and to obtain statistics on both inter-rater and intra-rater reliability. Examiners whose performance fluctuates or who mark consistently higher or lower than the group as a whole can be identified. This may help to identify scripts which ought to be marked a third time, or examiners who need further training or who should not be invited to mark this particular kind of examination in future.

b) TESTS OF SPEAKING

How different is marking speaking tests from marking written tests?

Many of the points made on marking tests of free writing also apply to speaking tests. Some differences are:

i) Interaction with candidate

The personality, behaviour and general competence of the examiner is even more important in tests which involve live interviews. This is the only kind of language test in which, unless audio or video tapes are used, assessment takes place simultaneously with the testing event. Often, an oral examiner has simultaneously to set up a rapport with the candidate (while at the same time controlling his/her own input) to take the candidate through the sequence of tasks set, and to make an assessment based on a rating scale, all within a time limit which may be as short as ten minutes per candidate.

ii) Timing and format of training

Training for oral examiners needs to take place before the speaking test. By the time they begin examining they must familiar with the assessment procedures and materials they are to use and with the rating scales they will mark on. It is very important to select and train competent examiners. It is useful for testing organisations to draw up a profile of what is required in an oral examiner.

Examiner training is often carried out using videos of a number of actual or simulated oral tests, in which potential candidates are 'examined' by experienced examiners. Trainee examiners watch the video and assign marks, which are then discussed, so that they can reach agreement on what score should be given.

What are the stages in training oral examiners?

Example speaking test: a test in English at advanced level (C1).
Part 1: interview (5 minutes)
Part 2: paired exercise, each candidate makes a presentation, asks and answers questions related to this, and takes part in a discussion (15 minutes)

i) Pre-training information

Trainee examiners are sent an information pack to study. The pack gives details of the form the speaking test takes and the objectives of the various parts of the test, administrative details, advice to candidates, etc.

ii) Group training sessions

Trainee examiners then take part in two half-day group training sessions.
The **first session** deals with the Part 1 interview:

- the format and procedures involved are discussed
- worksheets are completed to check trainee's understanding of the system
- assessment scales and scale descriptors are presented
- six interviews are shown on video and the trainees practise making an assessment
- the marks and issues are discussed and notes are provided to explain how the scores assigned to the candidates were arrived at.

The **second session** deals with administrative procedures and Part 2 of the test:

- three pairs of volunteer candidates are shown on video and the trainees practise making an assessment
- the marks and issues are discussed and notes are provided to explain how the scores assigned to the candidates were arrived at.

It may be possible for volunteer candidates to come to the training session and be 'examined' by the trainee examiners. This gives examiner and candidate a practical understanding of the examining process including timing, seating arrangements, dealing with materials, entering marks on the mark sheet etc.

iii) Standardisation

Once examiners have been trained, they must also go through a standardisation process on a regular basis to ensure that they are marking consistently.

This might consist of the following:

- a standardisation video of sample speaking tests is produced and marks agreed and notes drawn up by senior examiners
- examiners watch the video and work through a set of exercises assessing the candidates
- they discuss their assessments and compare with the marks / notes from the standardisation video, adjusting where necessary
- one final sample test is watched on the video and the senior examiner in charge of the meeting collects the marks to check subsequently (and provide feedback to any examiner(s) whose marking is not consistent with the level

Issues related to fairness in speaking test assessment

It is very important that oral examiners are competent and well trained. Some other factors affecting the fairness of assessment in speaking tests are:

i) Interlocutor frames

An interlocutor frame is a script which controls tightly the language used by oral examiners. It is a further way of standardising the testing process, and thus making it more consistent, less subjective and so fairer to each candidate. Sample tests on video for training purposes should demonstrate an effective use of the interlocutor frame with, for example, examiners who know the script well and do not need to read continuously from it.

If an interlocutor frame is not to be used, examiners must be trained to be aware of the need to control their use of language, always keeping in mind that their aim is to elicit the best possible performance from the candidate. Practice interviews which are videoed (or recorded) and discussed at training sessions can help examiners to recognise areas of their own performance which they need to improve; for example, using language above the level of the test, asking closed questions, talking too much etc.

Interlocutor frames are sometimes seen as barriers to ‘real communication’ between candidate and examiner since the examiner cannot engage in spontaneous response to what the candidate says. This has to be balanced against the issue of fairness, particularly in widely-taken international examinations where the experience of every candidate in every country should be as similar as possible with regard to the format of the test.

ii) Use of rating scales

The rating scales used must be finely-tuned enough to make useful distinctions between levels of performance, yet not too detailed to be practical in a situation where decisions have to be made very quickly. The scales must be expressed in terms which are explicit and unambiguous, and refer to aspects of the candidate's performance which are easily observable.

Examiners must also be aware of phenomena such as the 'halo effect', by means of which good performance in one area, such as pronunciation, tends to lead to over-generous assessment of weaker areas of performance. They must also avoid any tendency to judge candidates in relation to those seen immediately before them rather than in relation to the rating scales.

An example of a system of rating oral test candidates is given below. Marks are awarded on six scales, which together make up the candidate's profile of oral performance. On each scale, a mark of 0-5 is given for performance, ranging from relative naturalness and acceptability (5) to unintelligibility (0). There is no pass mark for the individual scales, which are put together to give a raw score out of 30.

1. FLUENCY Speed and rhythm, choice of structures, general naturalness and clarity.	**3. PRONUNCIATION** (sentences) Stress timing, rhythm and intonation patterns, linking of phrases.	**5. INTERACTIVE COMMUNICATION** Flexibility and linguistic resource in exchange of information and social interaction.
2. GRAMMATICAL ACCURACY Control of structures including tenses, prepositions, etc., to an effective level of communication.	**4. PRONUNCIATION** (individual sounds) Correct use of consonants and vowels in stressed and unstressed position for ease of understanding.	**6. VOCABULARY RESOURCE** Variety and correctness of vocabulary in the communicative context.

The rating scales used in the test of spoken English referred to above are given overleaf.

Rating	Grammar	Vocabulary	Pronunciation	Organization	Communication strategies and interaction	Interlocutor support	Task achievement
A	Wide range of structures including complex structures used accurately; very few errors	Extensive range; accurate and appropriate use; wide topic range; little hesitation in selection	Accurate and consistent use of all aspects of pronunciation.	Excellent - logically developed discourse; precise use of cohesive features.	Intended meaning communicated in all contexts; interaction initiated and maintained; topic changes responded to with ease.	Not required.	Task fully achieved and communicated successfully.
B	Full range of basic structures used with few errors; errors with complex structures.	Large range of everyday vocabulary though not always sufficient for discussion; little hesitation in selection.	Broadly accurate and consistent use of most aspects of pronunciation; foreign influence doesn't interfere.	Well organized - main points distinguished and appropriately sequenced; most discourse relationships well marked.	Intended meaning communicated in most contexts; interaction initiated and maintained; occasional difficulty in responding to changes of topic.	Occasionally required	Task is achieved but one or more of task requirements lacking.
C	Accurate use of basic structures; inaccurate use of complex structures causing occasional misunderstandings.	Moderate range for everyday use; some hesitation in selection.	Occasional inaccuracies but still intelligible; noticeable foreign accent but comprehensible.	Limited effectiveness; some discourse relationships inadequately marked; occasional need for clarification and repetition.	Main ideas communicated; repair strategies used; some difficulty in initiating interaction and responding to shifts of topic.	Frequently required.	Task only partly achieved; several task requirements are not fulfilled.
D	Inaccurate use of many basic structures; rare and inaccurate use of complex structures; frequent problems with intelligibility.	Restricted range sufficient for basic communication only; much hesitation in selection.	Inaccuracies sometimes result in unintelligible utterances; frequent serious errors interfere with communication.	Badly organized - discourse relationships not marked; frequent inappropriate sequencing; only basic discourse or conversational routines.	Main idea communicated in limited contexts; repair strategies rarely used; interaction seldom initiated; difficulty in responding to shifts of topic.	Continually required.	Task not achieved; task requirements not fulfilled.
E	Inaccurate use of most basic structures; no use of complex structures; largely unintelligible.	A few words and phrases only, usually inadequate for communication.	Inaccurate and inconsistent. Largely unintelligible.	No evidence of extended discourse or conversational rules; impossible to follow.	Great difficulty in communicating; unable to use repair strategies; needs others to maintain interaction.		

iii) Testing materials

A further factor in providing fair assessment in a speaking test involves providing the examiner with a sufficient variety of materials for a quick decision to be made at the beginning of an interview on what will be the best choice to use with that particular candidate. If a candidate expresses lack of interest in sport in the opening part of the test, the examiner may well decide not to use a set of materials focused on sport for the subsequent part(s) of the test.

The item writer needs to be aware of the need for flexibility in speaking test materials and for tasks to be accessible by candidates of different ages and backgrounds.

iii) Number and roles of examiners

The number of people involved in assessing can also affect the fairness of the assessment.

- interviews may be conducted on a one-to-one basis with the examiner conducting the interview and making the assessment
- there may be two examiners for each candidate, each examiner making an independent assessment
- there may be two examiners for two or three candidates at a time, the examiners taking the roles of interlocutor and assessor

If the roles of assessor and interlocutor can be separated, the assessor can concentrate on assigning a score, using analytical scales, while the interlocutor manages the interview and perhaps awarding a global impression mark to each candidate. In this way two independent marks are recorded.

Wherever it is possible for two examiners to be involved in assessing a candidate this has a clear advantage over one-to-one examining, both in terms of objectivity, and in making procedures to ensure reliability easier to apply.

If interviews are also recorded, it is easier to arrange for more than one examiner to assess at least those candidates over whose scores there is any uncertainty.

iv) Number and relationships of candidates

Candidates may be interviewed individually, or in pairs or larger groups. A further issue concerned with fairness is whether pairs or groups should be made up of candidates who already know each other, or whether strangers should be grouped together. Candidates who know each other may feel less nervous but may have ‘rehearsed’ some of their answers to a point beyond natural expression. Candidates who have not met before may feel uneasy at first but their interaction may be more communicative as they need to listen more actively to what their partner says.

Problems of fairness also arise from the effect of one personality on another, for example, a dominant and talkative candidate may limit another candidate's chances of expression. This issue needs to be addressed in examiner training and the speaking test criteria need to be in the public domain and easily available to teachers and candidates.

APPENDIX A

FORMULAS FOR CHECKING TEST RELIABILITY

Formula for:

Cronbach's alpha (coefficient alpha)

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum s_i^2}{s^2}\right]$$

Formula for:

Kuder-Richardson 20 (K-R20)

$$r = \frac{k}{k-1}\left[1 - \frac{\sum pq}{s^2}\right]$$

APPENDIX B

AN EXAMPLE OF A USE OF ENGLISH SUB-TEST WITH MARK SCHEMES

FCE Paper 3 (0102) June 2004

The Use of English sub-test which follows is composed of five item-based tasks. The item-based tasks are followed by mark schemes for this test.

Part 1 is marked by computer.
Parts 2, 3, 4 and 5 are marked by clerical markers.

Guidance is given to the clerical markers on:

- which answers are acceptable
- whether contracted forms are acceptable
- whether correct spelling is demanded
- the number of marks for each item

The clerical marker has to follow these instructions strictly, and refer to a supervisor or team leader in case of doubt.

SAMPLE PAPER: PAPER 3 USE OF ENGLISH

PAPER 3 Use of English

Part 1

For questions **1-15***, read the text below and decide which answer (***A, B, C** *or* **D***) best fits each space.*
*There is an example at the beginning (***0***).*

Mark your answers **on the separate answer sheet.**

Example:

0 **A** *consider* **B** *know* **C** *call* **D** *label*

0	A	B	C	D
			■	

SHOPPING MALLS

Victor Gruen, an American architect, revolutionised shopping in the 1950s by creating the type of shopping centre that we now (**0**) ..... a shopping mall.

Gruen's (**1**) ..... was to provide a pleasant shopping environment in the suburbs. This (**2**) ..... shutting out the noise of the city environment and also enabling people to shop in all kinds of weather. He (**3**) ..... on using building designs that he knew people would feel (**4**) ..... with, but placed them in landscaped 'streets' that were entirely enclosed and often covered with a curved glass roof. This was done to (**5**) ..... some of the older shopping arcades of city centres, but (**6**) ..... these housed only small speciality shops, Gruen's shopping malls were on a much grander (**7**) ..... .

Access to the whole shopping mall was gained by using the main doors, which (**8**) ..... the
Shopping 'streets' from the parking (**9**) ..... outside. As there was no need to (**10**) ..... out bad weather, shops no longer needed windows and doors, and people could wander (**11**) ..... from shop to shop. The space(**12**) ..... to build a shopping mall and its vast car parks can usually only be found in the suburbs or on the (**13**) ..... of the city. In many cities, shopping malls now (**14**) ..... much more than just shops; cinemas, restaurants and other forms of entertainment are also (**15**) ..... in popularity.

1	**A** direction	**B** aim	**C** search	**D** view
2	**A** resulted	**B** sought	**C** intended	**D** meant
3	**A** insisted	**B** demanded	**C** requested	**D** emphasised
4	**A** favourable	**B** agreeable	**C** comfortable	**D** enviable
5	**A** model	**B** imitate	**C** repeat	**D** shadow
6	**A** while	**B** even as	**C** besides	**D** in spite of
7	**A** measure	**B** height	**C** size	**D** scale
8	**A** disconnected	**B** withdrew	**C** separated	**D** parted
9	**A** strips	**B** lines	**C** areas	**D** plots
10	**A** hold	**B** get	**C** stay	**D** keep
11	**A** freely	**B** loosely	**C** simply	**D** entirely
12	**A** obliged	**B** required	**C** desired	**D** expected
13	**A** side	**B** limit	**C** edge	**D** extent
14	**A** contain	**B** concern	**C** consist	**D** compose
15	**A** becoming	**B** growing	**C** raising	**D** advancing

Part 2
For questions **16-30***, read the text below and think of the word which best fits each space. Use only* **one** *word in each space. There is an example at the beginning (***0***).*

Write your answers **on the separate answer sheet.**

Example:

0	among

DICTIONARIES

Dictionaries are (**0**) ..... the most important tools of self-education. (**16**) ..... Samuel Johnson wrote his influential English dictionary in the eighteenth century, the work kept him busy for seven years. At the end of that period, he (**17**) ..... written the meanings of over forty thousand words. Most modern dictionaries require a (**18**) ..... deal less time and effort to write because writers often use earlier dictionaries (**19**) ..... a source of reference.

(**20**) ..... it is possible for one person to write a dictionary alone, most dictionaries are team efforts. First of all, the writers, or lexicographers, draw up the rules that will guide their writing. For example, if a word has two meanings, they (**21**) ..... to agree about which order to put them (**22**) ..... . However, for much of the time, team members are able to work independently of (**23**) ..... other, on different parts of the dictionary.

(**24**) ..... one time, the starting point for deciding on those words to include used to be the
lexicographer's own knowledge. These days, some teams (**25**) ..... use of a large collection of examples of (**26**) ..... only writing but also everyday speech, which is known as a *corpus*. Teams also refer (**27**) ..... books and articles about language as (**28**) ..... as asking experts in particular subjects about the more specialised words. Finally, ordinary people are asked to say what they think about the (**29**) ..... the words are defined and (**30**) ..... they find the examples provided helpful or not.

Part 3

*For questions **31-40**, complete the second sentence so that it has a similar meaning to the first sentence, using the word given. **Do not change the word given**. You must use between **two** and **five** words, including the word given. Here is an example (**0**).*

***Example**:*

***0** You must do exactly what the manager tells you.*

Carry

You must ..... instructions exactly.

The space can be filled by the words 'carry out the manager's', so you write:

0	carry out the manager's

*Write **only** the missing words **on the separate answer sheet.***

---

31 Today's meeting is postponed and it will be held next week.

put

Today's meeting has ..... until next week.

32 Unfortunately, Kim couldn't go to the cinema because she didn't have any money.

able

If Kim had had some money, she ..... go to the cinema.

33 According to the report, the driver of the car was a policeman.

being

According to the report, the ..... by a policeman.

34 Nobody spoke for about five minutes.

before

It was about five minutes ..... anything.

35 Mr Johnson continued to get up at 6.30 even after he retired.

carried

Mr Johnson ..... at 6.30 even after he retired.

36 I prefer eating sandwiches to a cooked lunch.

rather

I ..... sandwiches than a cooked lunch.

37 ‘I’m sorry I behaved so badly,’ said George.

apologised

George ..... so badly.

38 There’s no chance of Jenny getting here on time.

possible

It won’t be ..... here on time.

39 ‘We really don’t need to leave early,’ said Elena.

point

‘There’s really ..... early,’ said Elena.

40 Cars couldn’t get onto the motorway because of an accident.

prevented

An accident ..... onto the motorway.

Part 4

For questions **41-55***, read the text below and look carefully at each line. Some of the lines are correct, and some have a word which should not be there.*
If a line is correct, put a tick (√) by the number **on the separate answer sheet***. If a line has a word which should* not *be there, write the word* **on the separate answer sheet***. There are two examples at the beginning (***0** *and* **00***).*

Examples:

0	√
00	more

MY GRANDMOTHER

0 The person I am going to write about is my grandmother. She is a
00 lively lady of more seventy-five years, although to look at her you would
41 think she is ten years younger aged. Now she lives in the city
42 and my brothers and I are often go to visit her at weekends. She
43 always gives to us coffee and some of her delicious home-made cake
44 and then we sit and listen to her so fascinating stories of the days
45 when she was being a girl. At that time, she lived deep in the
46 countryside in a place where there were very few cars or buses
47 and she was used to have to walk ten kilometres just to get to
48 school. There were no televisions, so she would amused herself
49 by playing outside with her friends until that it got dark. In
50 winter, when it was too completely dark, wet or cold to go outside,
51 she read books that she had borrowed from the local library and
52 learned a lot about places beyond the village she lived in there.
53 Eventually, by studying hard, she became a teacher and started the
54 work in a school in the city, and it was there that she met my
55 grandfather, who worked as an engineer in a large factory nearby it.

Part 5
For questions **56-65***, read the text below. Use the word given in capitals at the end of each line to form a word that fits in the space in the* **same** *line. There is an example at the beginning (***0***).*

Write your answers **on the separate answer sheet.**

Example:

0	tropical

ISLAND IN THE SUN

With its (**0**) ..... sunshine and warm welcome, this island is	**TROPIC**
hard to beat as a holiday destination. The (**56**) ..... west coast is the	**DELIGHT**
perfect (**57**) ..... for lovers of water sports and sunbathing. There is	**CHOOSE**
also an (**58**) ..... selection of restaurants, where the local seafood	**IMPRESS**
is (**59**) ..... recommended, and beaches of fine white sand face on to the	**HIGH**
calm Caribbean Sea.	

Other (**60**) ..... on the island include underwater trips in a submarine and	**ATTRACT**
a jazz festival held (**61**) ..... , early in January. It is also worth travelling	**ANNUAL**
along the wild east coast, which is often (**62**) ..... as it faces the Atlantic	**STORM**
Ocean. This makes the coast (**63**) ..... for swimming, unlike the calmer	**SUITABLE**
beaches on the west coast. Car hire is (**64**) ..... arranged here, and	**EASY**
there is a good road system, with a very (**65**) ..... bus service to take	**RELY**
you around the island.	

ANSWER KEY: Paper 3: Use of English

Part 1		Part 2		Part 3	
1	B	16	When	31	been | **put** off
2	D	17	had	32	would have been | **able** to
3	A	18	great	33	car was **being** | driven
4	C	19	as	34	**before** anybody | said
5	B	20	(Al)though	35	**carried** on | getting up
6	A	21	have	36	would **rather** | have
7	D	22	in	37	**apologised** for | behaving
8	C	23	each	38	**possible** for | Jenny to get
9	C	24	At	39	no **point** in | us/our leaving
10	D	25	make	40	**prevented** cars | from getting
11	A	26	not		
12	B	27	to		
13	C	28	well		
14	A	29	way		
15	B	30	whether		

APPENDIX C

AN EXAMPLE OF A COMPOSITION SUB-TEST WITH SAMPLE SCRIPTS AND COMMENTS

The writing sub-test which follows is marked by trained examiners, using the kind of mark schemes (General and Task Specific) explained on pages 189-192.

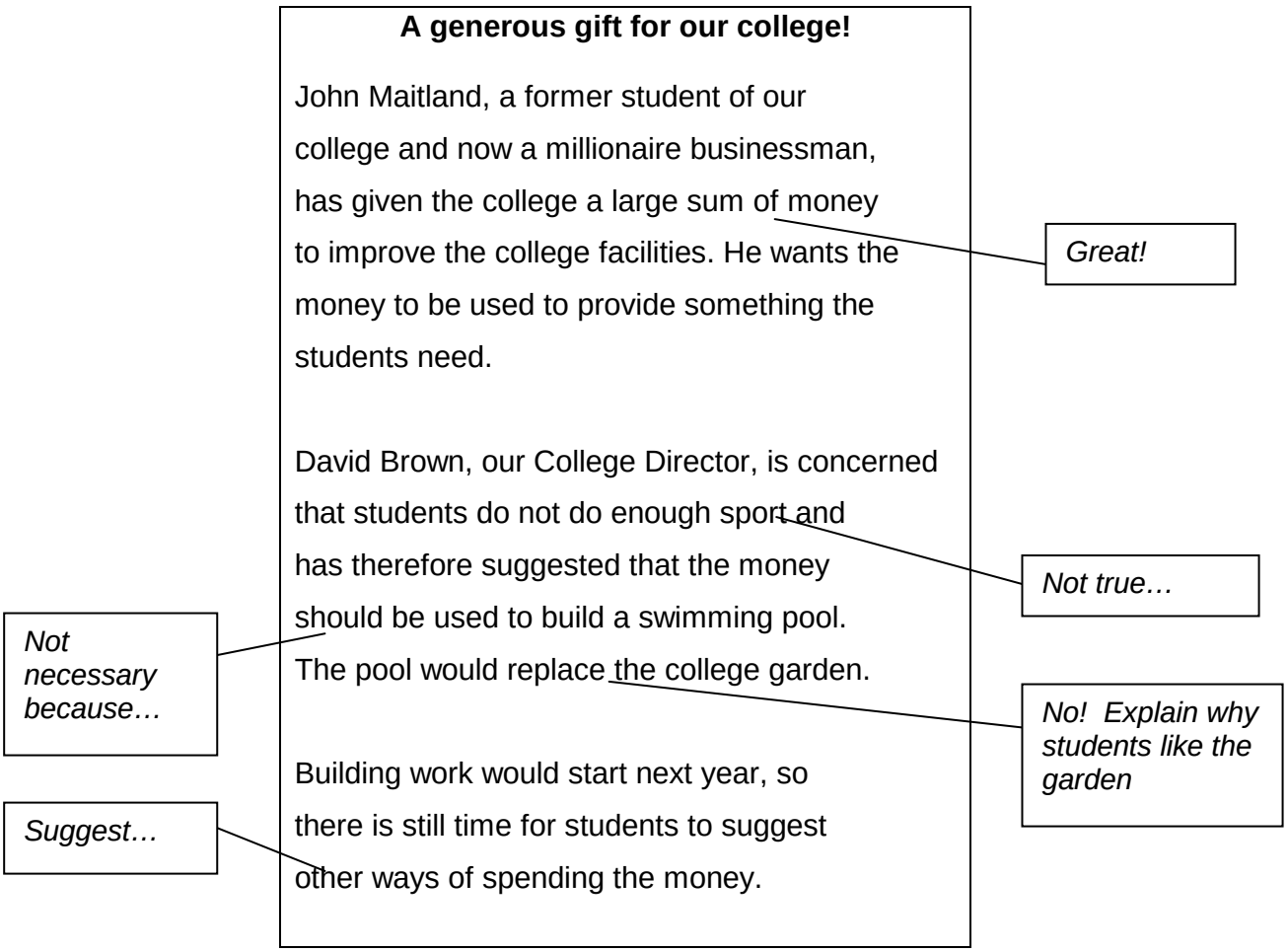
The sample scripts shown here are chosen by the chief examiner as exemplifying performance at certain levels. The training and standardisation process for this writing paper is explained on pages 169-172.

SAMPLE PAPER: PAPER TWO COMPOSITION

Instructions to candidates
Answer Question 1 and **one** of the Questions 2-5.
Each question in this paper carries equal marks.

Part 1
You **must** answer this question.

1 You are a student at a college in England and you have seen the article below in the college newsletter. Read the article and the notes you have made. Then write a letter to David Brown, the College Director, using **all** your notes.



Write a **letter** of between **120** and **180** words in an appropriate style on the opposite page. Do not write any postal addresses.

Part 2
Write an answer to **one** *of the questions* **2-5** *in this part. Write your answer in* **120-180** *words in an appropriate style on the opposite page. Put the question number in the box at the top of page* 5*.*

---

2 A British TV company is thinking of making a film about life in your area and has asked you to give them some information. Write a report describing the advantages of living in your area and saying how the area might change in the future.

Write your **report.**

3 You see this announcement in an English language magazine*.*

COLOUR
• What is your favourite colour and why?
• Why is colour so important in our lives?
Write us an article answering these questions.
The best article will be published in the magazine.

Write your **article.**

4 You have had a class discussion about sport. Now your English teacher has asked you to write a composition discussing the following question:

Should professional footballers be paid more than doctors?

Write your **composition.**

5 Answer **one** of the following two questions based on your reading of **one** of these set books. Write the letter **(a)** or **(b)** as well as the number **5** in the question box, and the **title** of the book next to the box. Your answer **MUST** be about one of the books below.

Round the World in Eighty Days – Jules Verne
Pride and Prejudice – Jane Austen
The Prisoner of Zenda – Anthony Hope
Deadlock – Sara Paretsky
Ghost Stories – retold by Rosemary Border

Either (a) An international magazine is looking for articles about brave characters in books.Write an **article** describing a character and explaining why you think that character shows courage in the book or short story you have read.

Or (b) 'At the end of a story, all the reader's questions are answered.' How true is this of the book or short story you have read? Write a **composition** explaining your views with reference to the book or short story you have read.

FIRST CERTIFICATE IN ENGLISH: PAPER 2 ASSESSMENT

There are two Parts to the writing paper, which carry equal marks. Part 1 is a compulsory task, where candidates must produce a 'transactional' letter, based on given input. In Part 2, candidates answer one task from a choice of four questions. Candidates should attempt **two** tasks only. If Part1 is not attempted, a candidate scores 0 for this question. In cases of more than one answer for Part 2, all answers are marked, and the highest mark taken.

ASSESSMENT FOCUS

The test focus in the two Parts is different, as outlined below. Examiners use the **General Mark Scheme** for both Parts, but also refer to a **Task-specific Mark Scheme** for each question.

PART 1 In assessing the transactional letter, the focus is on coverage and organisation of content, accuracy of language, and appropriacy of register and format to the target audience. Range will be defined by the parameters of the task. Some re-use of key words from the input is acceptable, but 'lifting' of phrases is penalised.

PART 2 The task types and topics in Part 2 offer the candidate more scope for linguistic initiative. In assessing a Part 2 task, the focus is on range of vocabulary and structure and on accuracy. There is greater flexibility in content and in the interpretation of the target audience. Articles or reports do not demand a single prescribed format, but credit is given for appropriate features of presentation.

ASSESSMENT PROCEDURE

Each piece of writing is assigned to a band between 0 and 5, as described in the General Mark Scheme, and awarded up to 3 points within the band, for example 3.1 / 3.2 / 3.3. Candidates are penalised for dealing inadequately with the requirements of the task specific markscheme.

LENGTH

120 - 180 words are asked for.

Below-length answers (50 - 100 words): assessment is confined to Bands 1 and 2. Answers containing fewer than 50 words receive 0.
Over-length answers (more than 200 words): the whole answer is assessed at first reading. At the second reading, a line is drawn across the page at the approximate place where the correct length is reached and close language assessment is confined to what comes above this line.

SPELLING AND PUNCTUATION

These are aspects of **accuracy** and are taken into account according to the extent to which they affect or impede communication. American usage and spelling are acceptable, provided that their use is consistent.

PARAGRAPHING

This is a function of organisation and format. For further guidance refer to Task Specific Mark Scheme.

HANDWRITING

If handwriting interferes with communication without preventing it, the assessment is brought down by one band. Total illegibility receives 0.

GENERAL MARK SCHEME

5	Full realisation of the task set. • All content points included with appropriate expansion. • Wide range of structure and vocabulary within the task set. • Minimal errors, perhaps due to ambition; well-developed control of language. • Ideas effectively organised, with a variety of linking devices. • Register and format consistently appropriate to purpose and audience. Fully achieves the desired effect on the target reader.
4	Good realisation of the task set. • All major content points included; possibly one or two minor omissions. • Good range of structure and vocabulary within the task set. • Generally accurate, errors occur mainly when attempting more complex language. • Ideas clearly organised, with suitable linking devices. • Register and format on the whole appropriate to purpose and audience. Achieves the desired effect on the target reader.
3	Reasonable achievement of the task set. • All major content points included; some minor omissions. • Adequate range of structure and vocabulary, which fulfils the requirements of the task. • A number of errors may be present, but they do not impede communication. • Ideas adequately organised, with simple linking devices. • Reasonable, if not always successful attempt at register and format appropriate to purpose and audience. Achieves, on the whole, the desired effect on the target reader.
2	Task set attempted but not adequately achieved. • Some major content points inadequately covered or omitted, and/or some irrelevant material. • Limited range of structure and vocabulary. • A number of errors, which distract the reader and may obscure communication at times. • Ideas inadequately organised; linking devices rarely used. • Unsuccessful/inconsistent attempts at appropriate register and format. Message not clearly communicated to the target reader.
1	Poor attempt at the task set. • Notable content omissions and/or considerable irrelevance, possibly due to misinterpretation of task set. • Narrow range of vocabulary and structure. • Frequent errors which obscure communication; little evidence of language control. • Lack of organisation, or linking devices. • Little or no awareness of appropriate register and format. Very negative effect on the target reader.
0	Achieves nothing: too little language for assessment (fewer than 50 words) or totally irrelevant or totally illegible.

All of these comments should be interpreted at FCE Level, and referred to in conjunction with a Task-specific Mark Scheme.

A maximum of 3 points can be awarded within each of Bands 1 - 5. 3.1 represents 'acceptable' performance at FCE. 5.3 may not be a flawless answer and should be awarded for top performance at FCE Level (Common Scale *Independent User*).

FCE 2 TASK-SPECIFIC MARK SCHEMES
SATISFACTORY BAND DESCRIPTORS (3.1 AND ABOVE)

QUESTION 1

CONTENT
Letter must include all the points in the notes:
1) expressive positive reaction to gift
2) point out that students do enough sport and/or have sports facilities
3) say that pool is not necessary/wanted
4) say that pool should not replace the garden and/or explain why students like the garden
5) suggest other way(s) of spending the money

ORGANISATION AND COHESION
Clear organisation of points, with suitable paragraphing and linking. Suitable opening and closing formulae.

APPROPRIACY OF REGISTER AND FORMAT
Formal or informal so long as consistent.

RANGE
Language for reacting positively, making suggestions, explaining/describing and disagreeing/commenting.

TARGET READER
Would be informed.

QUESTION 2

CONTENT
Report should describe the advantages of living in the writer's area and possible changes in the future. Not necessary to name the place.

RANGE
Language of description, opinion and explanation.

ORGANISATION AND COHESION
Report should be clearly organised with introduction and conclusion. Sub-headings an advantage.

APPROPRIACY OF REGISTER AND FORMAT
Register and range from neutral to formal but must be consistent throughout. Formal report layout not essential.

TARGET READER
Would be informed.

QUESTION 3 **CONTENT** Article should explain the writer's favourite colour(s) and why colour is important. **RANGE** Language of explaining, describing, giving opinions. **ORGANISATION AND COHESION** Clear organisation of ideas, with suitable paragraphing and linking. **APPROPRIACY OF REGISTER AND FORMAT** Any, so long as consistent. **TARGET READER** Would be informed.	**QUESTION 4** **CONTENT** Composition should discuss the question. Not necessary to answer the questions with a definite 'yes' or 'no'. **RANGE** Language of discussion and opinion. **ORGANISATION AND COHESION** Clear organisation of ideas, with suitable paragraphing and linking. **APPROPRIACY OF REGISTER AND FORMAT** Neutral composition. **TARGET READER** Would be informed.
QUESTION 5a **CONTENT** Writer should describe one or more character(s) in the book and explain how they show courage. **RANGE** Language of opinion, explanation and description. **ORGANISATION AND COHESION** Clear organisation of ideas, with suitable paragraphing and linking. **APPROPRIACY OF REGISTER AND FORMAT** Any, so long as consistent. **TARGET READER** Would be informed.	**QUESTION 5b** **CONTENT** Composition could agree or disagree with the statement. **RANGE** Language of opinion and explanation. **ORGANISATION AND COHESION** Clear organisation of ideas, with suitable paragraphing and linking. **APPROPRIACY OF REGISTER AND FORMAT** Neutral composition. **TARGET READER** Would be informed.

Question 1 **Script A**

Dear Mr Brown,

I am writing in response to the note in which you inform us about this fantastic news that the college will receive new funds.

I would like to react to your intention to spend this money on building a swimming pool. Indeed, after many conversations with other students, we unfortunately disagree with this idea because we feel that we practise enough sports with the equipments already provided such as the gymnasium, the football pitch and the tennis courts. Moreover, just five minutes walk from the college, there is an olympic swimming pool which offers very interesting fares for the students.

Therefore, I would like to point at the fact that students absolutely want to preserve 'their' garden because it allows us to have fresh air during our breaks, this is a place where we can relax. So we would like to preserve a little space of nature, which could be important for our concentration.

That is why I would like to suggest you to use the money on buying new computers, which we desperately need, or buying audio equipment for the Foreign Languages Department for instance.

I look forward to hear from.

Yours sincerely,

Question 1 **Script A**

CONTENT: all points covered, with good expansion

ACCURACY: minimal errors, due to ambition

RANGE: wide range

ORGANISATION AND COHESION: effectively organised

APPROPRIACY OF REGISTER AND FORMAT: consistently appropriate

TARGET READER: would be fully informed

Band 5

Question 2 **Script B**

About life in my area is Malaysia. I born in malaysia. In the malaysia the wherther is very hot and sunshine.

In my country got many nice place and the nice food in there. Seaside is very popular in my country. There very shinewy people like swimming and take boat go around the sea. In the night there walk in the beach. In my country have the night market. In the market many thing to sell. There have food. Clote, drink, CD movie, muzie and many thing there would many people eat in there.

In the future my country would change.

Question 2 **Script B**

CONTENT: poor attempt at task

ACCURACY: frequent errors obscure communication; lack of language control

RANGE: narrow range

ORGANISATION AND COHESION: attempt at organisation

APPROPRIACY OF REGISTER AND FORMAT: adequate

TARGET READER: a very negative effect on the target reader

Band 1

Question 3 **Script C**

COLOUR IS LIFE

To my opinion, nothing is more beautiful than a rainbow: it's 'the result of the fight between sun and rain'. In other words, colour is the way to express our feelings.

My favourite colour depends on my life, it changes as changes my life. But if I had to choose I would say blue. Blue because it reminds me the sea. I feel lively when I am by the sea, so I guess that being surrounded by blue helps me feeling lively.

But I have noticed that people, included me, have a tendancy in wearing bright colours when they are happy, and dark colours when they are not. Why?

I think colours can be qualified as a language. You can guess people personality by looking at their clothes, car, or else. But colours can influence our mood too. If we were living in a black and white world, people wouldn't be happy I think, everybody would be depressed. Colour gives life to anything. And there is no doubt that people feel better surrounded by colour.

For a happy world, we need colour in our lives, because colour is life.

Question 3 **Script C**

CONTENT: good development of task, leading to appropriate conclusion

ACCURACY: errors do not distract, although there are some careless slips

RANGE: good idiomatic structure, use of conditionals and range of vocabulary

ORGANISATION AND COHESION: well organised

APPROPRIACY OF REGISTER AND FORMAT: appropriate

TARGET READER: would be well informed

Band 4

Question 4 **Script D**

Enterteinment and Reality

Although I've been an apassionate of football since i was a child, I think that footballers should not be paid more than doctors.

Firstly, because I don't consider football as a profession. Football is only an enterteinment where footballers perfom. In few words, it's like actors and actresses in the movies. The doctors instead are saving lifes. What's more important or remarkable than that?
They are not valuated and also they are accussed of negligency. Sometimes. People should think about what is worth to be paid as deserves and not. I think doctors are real.

Question 4 **Script D**

CONTENT: task adequately achieved, although answer is slightly under-length

ACCURACY: a number of distracting and some impeding errors

RANGE: limited

ORGANISATION AND COHESION: poor

APPROPRIACY OF REGISTER AND FORMAT: inconsistent

TARGET READER: message not clearly communicated

Band 2

Question 5a **Script E**

'Round The World In Eighty Days'

One of the brave characters in book 'Round The World In Eighty Days' written by Jules Verne is Mr. Phileas Fogg. He makes a bet with his friends from The Reform Club that he can travel round the world in eighty days. He can loose everything if he doesn't do so. Don't you think it is a good start how to prove he's such a brave English gentleman? I bet you do.

It is a colourful race through Europe, Asia and America follows. They experience a lot of unexpected situations and delays. Mr Phileas Fogg and his French servant Passepartout face a danger travelling through India.

Let me tell you about one of the story I have read. As they travel through India meeting Indians priests who want to burn a beautyfull princess Aouda. They make a plan how to save her life. If Mr Fogg is not a brave man, he'll continue his journey without stoping and he wouldn't even think about saving Aouda. Of coure they were successful.

I can strongly recommended this book to anyone who is looking for a remarkable read.

Question 5a **Script E**

CONTENT: good realisation of task

ACCURACY: a number of non-impeding errors, particularly in the second half

RANGE: adequate

ORGANISATION AND COHESION: adequately organised

APPROPRIACY OF REGISTER AND FORMAT: reasonably appropriate

TARGET READER: achieves the desired effect

Band 3