Manual for Language Test Development and Examining

For use with the CEFR

Produced by ALTE on behalf of the
Language Policy Division, Council of Europe

ALTE Association of Language Testers in Europe

Languages in/for Education
Langues dans/pour l'Education
Language Policy Division
Division des Politiques linguistiques

COUNCIL OF EUROPE CONSEIL DE L'EUROPE

Manual for Language Test Development and Examining

For use with the CEFR

Produced by ALTE on behalf of the
Language Policy Division, Council of Europe

Language Policy Division
Council of Europe (Strasbourg)
www.coe.int/lang

Contents

Foreword

This Manual is a welcome and timely addition to the 'toolkit', which offers support for specific groups in the use of the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR). We are grateful to the Association of Language Testers in Europe (ALTE) which was commissioned to prepare this volume by the Council of Europe and which, in keeping with the spirit of the International Non-Governmental Organisation (INGO) participatory status which it enjoys with the Council of Europe, continues to make a significant contribution to the effective use of the CEFR.

The CEFR is intended to provide a shared basis for reflection and communication among the different partners in the field, including those involved in teacher education and in the elaboration of language syllabuses, curriculum guidelines, textbooks, examinations, etc., across the member states of the Council of Europe. It is offered to users as a descriptive tool that allows them to reflect on their decisions and practice, and situate and co-ordinate their efforts, as appropriate, for the benefit of language learners in their specific contexts. The CEFR is therefore a flexible tool to be adapted to the specific contexts of use – a fundamental aspect fully reflected in the level system itself which can be adapted and exploited flexibly for the development of learning/teaching objectives and for assessment, as in the development of Reference Level Descriptions (RLDs) for particular languages and contexts.

The illustrative descriptors, based on those 'which have been found transparent, useful and relevant by groups of native and non-native teachers from a variety of educational sectors with very different profiles in terms of linguistic training and teaching experience' (CEFR, p.30), are not intended to be totally comprehensive or in any way normative. Users are invited to use and adapt or supplement them according to context and need. This practical Manual provides valuable guidance for users in constructing proficiency tests in this spirit, related to the CEFR levels in a principled and non-prescriptive manner.

The need to ensure quality, coherence and transparency in language provision, and the increasing interest in the portability of qualifications, have aroused great interest in the CEFR levels which are used in Europe and beyond as a reference tool and a calibrating instrument. While welcoming this, we would also encourage users to explore and share experiences on how the CEFR in its various dimensions can be further exploited to support and acknowledge the lifelong development of the (uneven and dynamic) plurilingual profile of language learners who ultimately need to take responsibility for planning and assessing their learning in the light of their evolving needs and changing circumstances. The Council of Europe's initiatives to promote plurilingual and intercultural education, and a global approach to all languages in and for education, present new challenges for curriculum development, teaching and assessment, not least that of assessing learners' proficiency in using their plurilingual and intercultural repertoire. We look forward to the essential contribution of professional associations such as ALTE in our efforts to promote the values of the Council of Europe in the field of language education.

Joseph Sheils
Language Policy Division
Council of Europe

Introduction

Background

Since the Common European Framework of Reference for Languages (CEFR) was published in its finalised version in 2001, it has attracted an increasing amount of interest, not only in Europe, but also globally. The impact of the CEFR has exceeded expectation and there is no doubt that the Framework has helped to raise awareness of important issues related to the learning, teaching and assessment of languages. The Council of Europe has also encouraged the development of a 'toolkit' of resources to inform and facilitate the intended uses of the CEFR by policy makers, teachers, assessment providers and other interested parties.

As noted by Daniel Coste (2007), one of the authors of the CEFR, the influence on language assessment has been particularly noteworthy, and the processes involved in aligning language tests to the CEFR's Common Reference Levels have received more attention than other aspects of the Framework. A number of useful tools from the toolkit are now available to assessment providers and other practitioners with an interest in language testing. These include:

- the *Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment* (Council of Europe, 2009)
- a technical *Reference Supplement to the Manual for Relating Examinations to the CEFR* (Banerjee 2004; Verhelst 2004 a,b,c,d; Kaftandjieva 2004; Eckes 2009)
- exemplar materials illustrating the CEFR levels
- content analysis grids for speaking, writing, listening and reading materials
- developing Reference Level Descriptions for English and other languages.

The Council of Europe has also provided forums (Reflections on the use of the Draft Manual for Relating Language Examinations to the CEFR, Cambridge, 2007; pre-conference seminar, EALTA Conference, Athens, 2008) where practitioners shared their reflections on the use of the Manual and their experience in using the different linking stages suggested in it.

The Association of Language Testers in Europe (ALTE), as an International Non-Governmental Organisation (INGO) with consultative status in the Council of Europe, has contributed to the resources which make up the toolkit, including the EAQUALS/ALTE European Language Portfolio (ELP) and the ALTE content analysis grids. It was also represented by Dr Piet van Avermaet in the authoring team which produced the *Manual for Relating Language Examinations to the CEFR*. Together with the Council of Europe Language Policy Division, ALTE is keen to encourage users of the toolkit to make effective use of the CEFR in their own contexts to meet their own objectives.

The aim of this Manual

The *Manual for Relating Language Examinations to the CEFR* mentioned above was designed specifically to address the alignment of tests to the Framework and together with its Reference Supplement, the document sets out and discusses a general approach and a number of options which are available, including standard setting.

The *Manual for Language Test Development and Examining* is designed to be complementary to that Manual; it focuses on aspects of test development and examining which were not covered in the other Manual. It is, in fact, a revised version of an earlier Council of Europe document originally known as a *Users' Guide for Examiners* (1996), which was one of several Users' Guides commissioned by the Council of Europe to accompany the first draft of the CEFR in 1996/7.

ALTE was responsible for producing the original version. Over the past decade or so, developments in validity theory and the growing use and influence of the CEFR have highlighted the need for a thorough updating of the document. ALTE was pleased to have been invited again to coordinate these revisions in 2009/10 and many individual ALTE members and affiliates have contributed to the process.

In making the revisions, it has been useful to remind ourselves of the origins and purposes of the CEFR itself and to reflect this in the way that this new Manual is structured and targeted at potential users.

As a common framework of reference, the CEFR was primarily intended as 'a tool for reflection, communication and empowerment' (Trim, 2010). It was developed to facilitate common understanding in the fields of language learning, teaching and assessment and provides a compendious discussion of language education, providing a common language for talking about every aspect of it. It also provides a set of reference levels for identifying levels of language proficiency, from near-beginner (A1) to a very advanced level (C2), and over a range of different skills and areas of use.

These features make it an appropriate tool for comparison of practices across many different contexts in Europe and beyond. However, in fulfilling this purpose (as a common reference tool), it cannot be applied to all contexts without some kind of user intervention to adapt it to local contexts and objectives.

Indeed, the authors of the CEFR made this point very clearly; in the introductory notes for the user, for example, they state that 'We have NOT set out to tell practitioners what to do or how to do it' (p.xi), a point which is reiterated several times throughout the text. Subsequent resources making up the toolkit, such as the *Manual for Relating Language Examinations to the CEFR*, have followed suit. The authors of that Manual clearly state that it is not the only guide to linking a test to the CEFR and that no institution is obliged to undertake such linking (p.1).

At a Council of Europe Policy Forum on the use of the CEFR in Strasbourg in 2007, Coste noted how contextual uses which are seen as deliberate interventions in a given environment can take 'various forms, apply on different levels, have different aims, and involve different types of player.' He goes on to state: 'All of these many contextual applications are legitimate and meaningful but, just as the Framework itself offers a range of (as it were) built-in options, so some of the contextual applications exploit it more fully, while others extend or transcend it.' Thus, when considering issues of alignment, it is important to remember that the CEFR is not intended to be used prescriptively and that there can be no single 'best' way to account for the alignment of an examination within its own context and purpose of use.

As Jones and Saville (2009: 54-55) point out:

> '… some people speak of applying the CEFR to some context, as a hammer gets applied to a nail. We should speak rather of referring context to the CEFR. The transitivity is the other way round. The argument for an alignment is to be constructed, the basis of comparison to be established. It is the specific context which determines the final meaning of the claim. By engaging with the process in this way, we put the CEFR in its correct place as a point of reference, and also contribute to its future evolution.'

While the *Manual for Relating Language Examinations to the CEFR* focuses on 'procedures involved in the justification of a claim that a certain examination or test is linked to the CEFR' and 'does not provide a general guide how to construct good language tests or examinations' (p.2), the complementary approach adopted in this Manual starts from the test development process itself and shows how a link to the CEFR can be built into each step of this process, in order to:

- specify test content
- target specific language proficiency levels
- interpret performance on language tests in terms that relate to a world of language use beyond the test itself

This Manual, therefore, has a wider aim than the three main uses outlined in the CEFR itself, namely:

- the specification of the content of tests and examinations
- stating the criteria for the attainment of a learning objective, both in relation to the assessment of a particular spoken or written performance, and in relation to continuous teacher-, peer- or self-assessment
- describing the levels of proficiency in existing tests and examinations thus enabling comparisons to be made across different systems of qualifications.

(CEFR:19)

It aims to provide a coherent guide to test development in general which will be useful in developing tests for a range of purposes, and presents test development as a cycle, because success in one stage is linked to the work done in a previous stage. The whole cycle must be managed effectively in order for each step to work well. Section 1.5 offers an overview of the cycle, which is then elaborated in detail in each chapter:

Section 1 – introduces the fundamental concepts of language proficiency, VALIDITY, RELIABILITY and fairness

Section 2 – developing the test – goes from the decision to provide a test through to the elaboration of final SPECIFICATIONS

Section 3 – assembling tests – covers item writing and construction of tests

Section 4 – delivering tests – covers the ADMINISTRATION of tests, from registration of test takers through to the return of materials

Section 5 – marking, grading and reporting of results – completes the operational cycle

Section 6 – monitoring and review – how the cycle can be repeated over time in order to improve the quality and usefulness of the test.

The reader of this Manual

This Manual is for anyone interested in developing and using language tests which relate to the CEFR. It is written to be useful to novice language testers as well as more experienced users. That is, it introduces common principles which apply to language testing generally, whether the exam provider is a large organisation preparing tests for thousands of test takers in various locations, or a single teacher who wishes to test the students in his or her own classroom. The principles are the same for both high- and low-stakes tests, even though the practical steps taken will vary.

We assume that readers are already familiar with the CEFR, or will be ready to use it together with this Manual when developing and using tests.

How to use this Manual

Even though the principles of language testing introduced here have general applicability, the test provider must decide how to apply them in their own context. This Manual provides examples, advice and tips for how certain activities might be conducted. However, this practical advice is likely to be more relevant to some contexts than to others, depending on the purpose of the test and the resources available to develop it. This need not mean that this Manual is less useful for some readers: if users understand the principles, then they can use the examples to reflect on how to implement the principles in their own context.

Apart from the CEFR itself, many other useful resources also exist to help with relating language tests to the CEFR and this Manual is just one part of the toolkit of resources which have been developed and made available by the Council of Europe. For this reason, it does not attempt to provide information or theory where it is easily accessible elsewhere. In particular, as noted above, it tries not to duplicate information provided in the *Manual for Relating Language Examinations to the CEFR* but is complementary to it.

This Manual need not be read from cover to cover. If different tasks in test development and administration are to be performed by different people, each person can read the relevant parts. However, even for those specialising in one area of language testing, the entire Manual can provide a useful overview of the whole testing cycle.

Further reading at the end of each section points readers either towards resources covering the topics in greater depth, or towards practical tools. This is followed by key questions, which will help strengthen understanding of what has been read.

This Manual is a non-prescriptive document which seeks to highlight the main principles and approaches to test development and assessment which the user can refer to when developing tests within their own contexts of use. It is not a recipe book for developing test questions based on the CEFR's illustrative scales since although the six reference levels of the CEFR are clear and detailed enough to provide a common reference tool, the scales were not designed to provide the basis for precise equating.

Indeed, in one of the early drafts of the framework document (Strasbourg 1998), the illustrative scales were included in the appendix as examples and did not appear in the body of the text. The only scales to be included in the main text were the Common Reference Levels. The original layout of the text in the 1998 draft reinforced the different statuses and functions of the general *reference levels* and the more specific *illustrative scales*. This approach underlined the tentative nature of the scales, some of which were uncalibrated and were under-represented at the C-levels.

In the 1998 draft version of the CEFR, the tentative status of the illustrative scales was made explicit in the text (p. 131):

> 'The establishment of a set of common reference points in no way limits how different sectors in different pedagogic cultures may choose to organise or describe their system of levels and modules. It is also to be expected that the precise formulation of the set of common reference points, the wording of the descriptors, will develop over time as the experience of member states and of institutions with related expertise is incorporated into the description.'

The risk of using the scales in an overly prescriptive way is that this might imply a 'one-size-fits-all' approach to measuring language ability. However, the functional and linguistic scales are intended to illustrate the broad nature of the levels rather than define them precisely. Thus, given the many variations in demographics, contexts, purposes, and teaching and learning styles, it is not possible to characterize a 'typical B1' student, for example. As a corollary, this makes it difficult to create a syllabus or test for B1, or indeed any other level, which is suitable for all contexts.

In order for the CEFR to have a lasting and positive impact, its principles and practices need to be integrated into the routine procedures of assessment providers. This will enable alignment arguments to be built up over time as the professional systems develop to support the claims being made, and involves working with the text of the CEFR as a whole and adapting it where necessary to suit specific contexts and applications.

Since it is not possible to ensure stable and consistent evidence of alignment from a single standard-setting exercise, it is important for examination providers to seek to provide multiple sources of evidence which have been accumulated over time. This means that the recommendations found in the *Manual for Relating Language Examinations to the CEFR* and other resources in the toolkit used for alignment purposes need to be integrated within the standard procedures of the assessment provider and should not be seen as purely 'one-off events'.

This is what this Manual encourages the reader to do by emphasising the importance of designing and maintaining systems which enable standards to be set and monitored over time.

Conventions used in this Manual

Throughout this Manual, the following conventions are observed:

- The *Manual for Language Test Development and Examining* is referred to as *this Manual*.
- The *Common European Framework of Reference for Languages: Learning, teaching, assessment* (CEFR) is referred to as the *CEFR*.
- The organisation responsible for producing the test is referred to as the *test provider*. Terms such as *test developer* are occasionally used to refer to those carrying out a specific function within the testing cycle.
- Words which are found in the Glossary (Appendix VIII) are highlighted by SMALL CAPS the first time they appear in this Manual and at other times where it is thought helpful to the reader.

Dr Michael Milanovic
ALTE Manager

1 Fundamental considerations

This Manual's practical guidelines for constructing language tests require a sound basis in some underlying principles and theory. This chapter considers the following issues:

- how to define language proficiency
- how to understand *validity* as the key property of a useful test
- how to understand *reliability*
- fairness in testing.

This section also gives an outline of the test development process covered by the subsequent chapters.

1.1 How to define language proficiency

1.1.1 Models of language use and competence

Language in use is a very complex phenomenon which calls on a large number of different skills or competences. It is important to start a testing project with an explicit model of these competences and how they relate to each other. Such a model need not represent a strong claim about how language competence is actually organised in our brains, although it might do; its role is to identify significant aspects of competence for our consideration. It is a starting point for deciding which aspects of use or competence can or should be tested, and helps to ensure that test results will be interpretable and useful. The mental characteristic identified by the model is also called a TRAIT or CONSTRUCT.

1.1.2 The CEFR model of language use

Influential models of language competence have been proposed by several authors (e.g. Bachman 1990, Canale and Swain 1981, Weir 2005).

For this Manual it makes sense to start with the CEFR, which proposes a general model that applies to language use and language learning. This 'ACTION-ORIENTED APPROACH' is introduced in one paragraph, describing:

> '… the actions performed by persons who as individuals and as social agents develop a range of ***competences***, both ***general*** and in particular ***communicative language competences***. They draw on the competences at their disposal in various contexts under various ***conditions*** and ***constraints*** to engage in ***language activities*** involving ***language processes*** to produce and/or receive **texts** in relation to ***themes*** in specific ***domains***, activating those ***strategies*** which seem most appropriate for carrying out the ***tasks*** to be accomplished. The monitoring of these actions by the participants leads to the reinforcement or modification of their competences' (CEFR:9 (emphasis in original)).

This paragraph identifies the major elements of the model, which are then presented in more detail in the text of the CEFR. In fact, the headings and subheadings in Chapters 4 and 5 of the CEFR can be seen as defining a hierarchical model of elements nested within larger elements.

Figure 1 illustrates this by showing some of the headings and subheadings of Chapter 5, *The user/learner's competences*. It shows competences sub-divided into two: *General competences* (such as *Declarative knowledge* and *Existential competence*, not shown here) and *Communicative language competences*, which are further subdivided into three: *Linguistic*, *Sociolinguistic* and *Pragmatic competences*. Each of these is further subdivided.

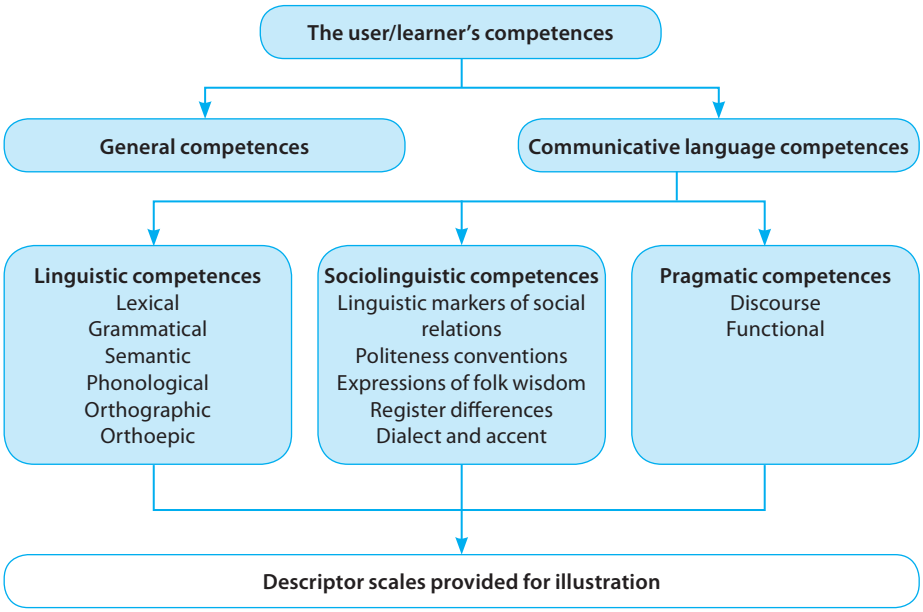


Figure 1 ***A partial view of CEFR Chapter 5: The user/learner's competences***

Similarly, Chapter 4 discusses the communicative purposes and ways in which language is used. As Figure 2 indicates, this involves consideration of *what* is communicated (themes, tasks and purposes) but also of the activities and strategies, and hence the functional language skills, which learners demonstrate when they communicate. For clarity, Figure 2 illustrates only a part of this complex hierarchy.

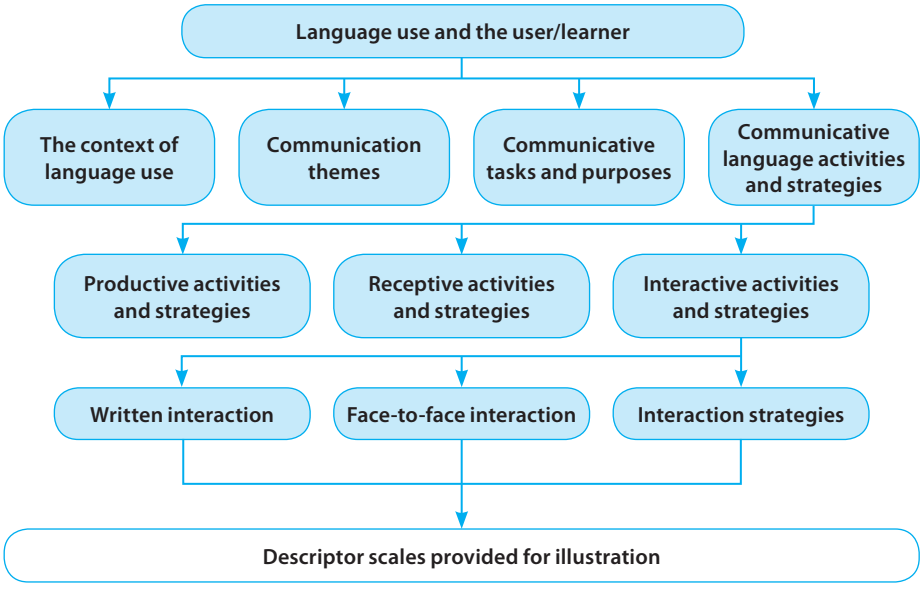


Figure 2 ***A partial view of CEFR Chapter 4: Language use and the language user/learner***

1.1.3 Operationalising the model

When considering how to operationalise the MODEL OF LANGUAGE USE, we must consider two important aspects which have a considerable influence over how the final test will look: AUTHENTICITY of ITEMS and TASKS and the discreteness with which competences are tested.

Authenticity

Two important aspects of authenticity in language testing are *situational* and *interactional* authenticity. *Situational* authenticity refers to the accuracy with which tasks and items represent language activities from real life. *Interactional* authenticity refers to the naturalness of the interaction between test taker and task and the mental processes which accompany it. A test task based on listening for specific information can be made more situationally authentic if an everyday context, such as a radio weather forecast, is created. It can be made more interactionally authentic if the test taker is given a reason for listening – e.g. planning a picnic that week and must select a suitable day.

In language tests, we often have to balance different aspects of authenticity to create an appropriate task. For example, it is necessary to adapt materials and activities to learners' current level of language proficiency in the target language. This tailoring means that while the materials may not be linguistically authentic, the situations that learners engage with, and their interaction with the texts and with each other, *can* be authentic.

To make an item or task more situationally authentic, the key features of the real-life task must be identified and replicated as far as possible. Greater interactional authenticity can be achieved in the following ways:

- using situations and tasks which are likely to be familiar and relevant to the intended test taker at the given level
- making clear the *purpose* for carrying out a particular task, together with the intended *audience*, by providing appropriate contextualisation
- making clear the *criterion for success* in completing the task.

Integrating competences

Competences can appear to be discrete when we define a model of language use. However, it is very difficult to clearly separate competences in authentic tasks. This is because any communicative act involves many competences at the same time. For example, when a language learner tries to understand someone who has stopped them on the street to ask for directions, a number of competences are involved: grammatical and textual competence to decode the message, sociolinguistic competence to understand the social context of the communication, and illocutionary competence to interpret what the speaker wishes to achieve.

When we are designing a test task, it is important to be clear about the balance between competences needed for a successful RESPONSE. Some competences will be more important than others – these will form the focus of the task. The task should elicit sufficient appropriate language for a judgement to be made about the test taker's ability in the chosen competence(s). The way the response is MARKED or rated is also important to consider (see Sections 2.5 and 5.1.3): ability in the chosen competence(s) should be the basis of the mark.

1.1.4 The Common Reference Levels of the CEFR

In addition to the model presented above, the CEFR also identifies a framework of six levels of communicative language ability as an aid to setting learning objectives, and measuring learning progress or proficiency level. This conceptual Framework is illustrated by a set of descriptor scales, expressed in the form of '*Can Do*' *statements*.

An example 'Can Do' statement for low-level (A1) reading is:

Can understand familiar names, words and very simple sentences, for example on notices and posters or in catalogues.

Compare this with a high-level (C2) descriptor:

Can read with ease virtually all forms of the written language, including abstract, structurally or linguistically complex texts such as manuals, specialised articles and literary works.

Council of Europe (2001:26–7)

The six proficiency levels are named as follows:

C2	Mastery	} Proficient user
C1	Effective Operational Proficiency	
B2	Vantage	} Independent user
B1	Threshold	
A2	Waystage	} Basic user
A1	Breakthrough	

As language testers we should understand the 'Can Do' statements correctly. They are:

- illustrative.

Therefore, they are not:

- exhaustive
- prescriptive
- a definition
- a curriculum
- a checklist.

The 'Can Do' statements offer guidance to educators so that they can recognise and talk about ability levels. We can use them as a guide for test development but should not feel that adopting them means the work of defining ability levels for the test has been completed.

Test developers must decide which 'Can Do' statements are most relevant to their context. For example, the DOMAIN of their test: for the teaching and testing of hotel staff, the 'Goal-orientated co-operation' descriptors (CEFR:4.4.3.1) may be useful, whereas those dealing with 'Watching TV and film' (CEFR:4.4.2.3) will probably not be. If the available illustrative SCALES or other materials in the CEFR toolkit do not match the context closely enough, then they can be supplemented with 'Can Do' statements from other sources, or new ones written for the context.

Aligning tests to the CEFR

Working in this way, we can see that the work of aligning a test to the CEFR starts with adapting the CEFR to the context of the test. This is because the CEFR sets out to be '*context-free* in order to accommodate generalisable results from different specific contexts' but at the same time '*context-relevant*, relatable to or translatable into each and every relevant context' (CEFR:21).

Aligning should not be an attempt to apply the CEFR to any context in the same rigid or mechanical way. The test developer must be able to justify the way they relate or 'translate' the CEFR to their context partly by explaining features of their context.

Other important contextual features include characteristics of the test takers. For example, there are important differences between learners, in terms of age and cognitive development, purpose in learning a language, and so on. Some of these differences actually define characteristically different groups of learners. Language tests are often designed with one such group of learners in mind, such as young learners or adults. Both groups may be related to the CEFR, but a B1 young learner and a B1 adult will demonstrate different kinds of 'B1-ness' because different descriptors apply.

Learners often differ in their profile of skills (some may be better listeners than readers, and others better readers than listeners). This makes it difficult to compare them on a single scale. For this reason, two test takers may be described as B1 due to different strengths and weaknesses. If it is important to distinguish between ability in different skill areas, skills may be tested separately and skill-specific descriptors used as the basis of defining skill-specific ability levels.

There is, however, an important limitation when adapting the CEFR to a particular context. The CEFR is only intended to describe language ability according to the model of language use described in 1.1.2 of this Manual. No attempt should be made to LINK knowledge or ability not covered in this model, such as an understanding of literature in a foreign language.

1.2 Validity

1.2.1 What is validity?

Validity can be simply defined: a test is valid if it measures what we intend it to measure. Thus for example, if our test is intended to measure communicative ability in Italian, and people score systematically higher or lower due to their Italian ability, then our test is valid. This rather narrow definition has been extended in recent years to encompass the way tests are *used*, i.e. validity relates to: 'the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests' (AERA, APA, NCME 1999).

This extended definition stresses the social IMPACT of tests, and the need for tests to provide adequate information for making what may be important decisions about individual test takers. In this view we cannot speak of a test being valid, in some absolute sense. Instead, validity relates to the way in which test results are used for a particular purpose: it is the interpretation of the meaning of test results for an individual test taker which may be valid or invalid.

Bachman (1990) refers this to the specific case of language by stating that tests should support inference to some domain of *target language use*. That is, in order to judge the validity of test results, we must first state what we expect a test taker to be able to do using language in the real world, and then decide whether the test provides good evidence of their ability to do so. The CEFR outlines a useful approach to define achievement in specific domains of use. Its illustrative descriptors offer a starting point.

1.2.2 Validity and the CEFR

When we report test results in terms of the CEFR, we are claiming to be able to interpret test performance in terms of our definition of test takers at particular CEFR levels. Validity comes down to demonstrating that what we claim is true: that a learner reported to be at B1 actually *is* at B1 according to the evidence we can provide.

The kind of evidence needed may vary depending on the testing context. The CEFR's model of language learning/use presented above can be called *socio-cognitive*: language is both an internalised set of competences, and an externalised set of social behaviours. Depending on the context, a language test may focus more on one than the other, and this affects the evidence for its validity:

- if the focus is more on use, then validity evidence will relate to language actually being used for various communicative purposes
- if the focus is more on competence, then validity evidence will relate to cognitive skills, strategies and language knowledge that support inference about potential ability for language use.

In the latter case it will be important to show that the test tasks engage the same skills, strategies and language knowledge that would be needed in the target language use domain – that is, that they have *interactional authenticity* (see 1.1.3).

Both kinds of evidence can support the CEFR-related validity of a language test. The balance between them depends on the requirements of the specific context. A language test for sales people would probably weight ability to use language very strongly; a language test for schoolchildren might put more stress on competence.

1.2.3 Validity within a test development cycle

Validity thus links performance on the tasks in a test to an inference about the test taker's language ability in a world beyond the test. Clearly, designing and constructing tasks is a critical step but other steps are also crucial.

In this section, validity is related to the cycle of producing tests (see 1.5), so that the influence of other test production stages can be seen. The stages of test production mean that the series of steps is described as sequential, and each step must be satisfactorily completed if the final inference is to be valid.

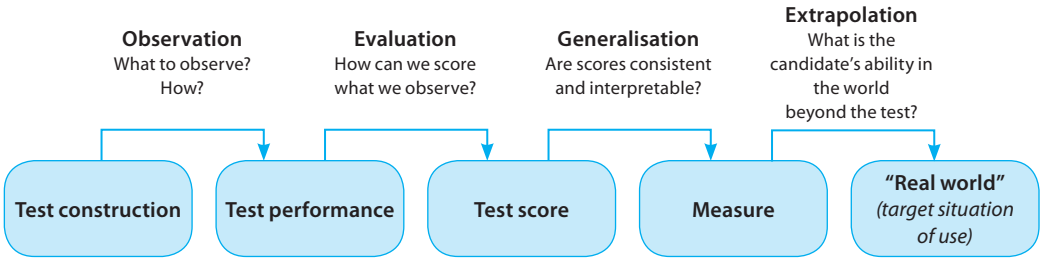


Figure 3 ***Chain of reasoning in a validity argument (adapted from Kane, Crooks and Cohen 1999; Bachman 2005)***

Figure 3 illustrates these steps in a schematic way:

1. The test is designed to elicit a sample of performance which is interpretable, based on a model of the learner's competences. For example, a test taker may be asked to write a letter to a friend on a particular topic.
2. The test performance is scored. Which features of performance will be rewarded, or penalised? In our example, these features will be related to communicative ability described by the model of language use, including REGISTER (sociolinguistic competence), lexical, grammatical and orthographical competence (linguistic competences), etc.
3. So far, the test scores are numbers which relate only to a single performance on a specific task. How can they be generalised – would the test taker get the same result on a different occasion, on a different test version? This question concerns reliability (see Section 1.3). A second aspect of generalisation concerns aligning to a wider proficiency scale, as one test form may be easier than another, for example, and we would wish to identify and compensate for that (see Appendix VII).
4. So far we have described performance in the world of the test, but we wish to extrapolate to the world beyond the test. This is where we would link a measure to a CEFR level as we describe what the test taker should be able to do in the real world, using 'Can Do' statements as a guide.
5. Based on this, we can make decisions about the test taker.

From this brief outline it is clear that validity, including a claimed link to the CEFR, depends on each step in the test construction and administration cycle. Validity is built into the whole process.

Appendix I provides guidance in constructing a validity argument.

1.3 Reliability

1.3.1 What is reliability?

Reliability in testing means consistency: a test with reliable scores produces the same or similar result on repeated use. This means that a test would always rank-order a group of test takers in nearly the same way. It does *not* mean that the same people would pass or fail, because the pass mark might be placed differently. The term *dependability* is generally used where we are interested in the consistency and accuracy of an exam grade or result.

Note that high reliability does not necessarily imply that a test is good or interpretations of the results are valid. A bad test can produce highly reliable scores. The opposite is not true, though: for the valid interpretation of test results, scores *must* have acceptable reliability, because without it the results can never be dependable or meaningful.

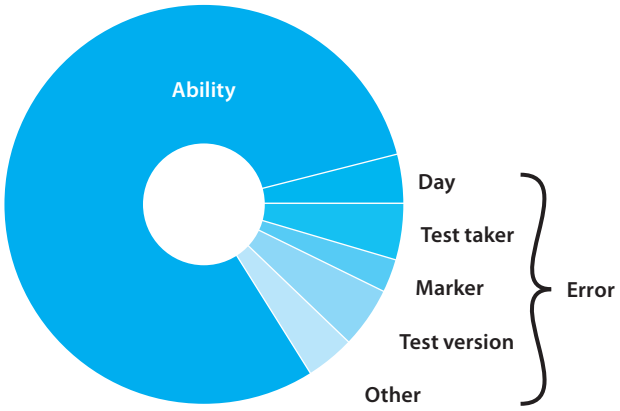


Figure 4 ***Some sources of error in a test score***

Test scores vary between test takers. Reliability is defined as the proportion of this test score variability which is caused by the ability measured, and not by other factors. The variability caused by other factors is called *ERROR*. Note that this use of the word error is different from its normal use, where it often implies that someone is guilty of negligence. All tests contain a degree of error.

Figure 4 shows some of the common sources of error:

- the day of the test session (the weather, the administration, etc., might be different)
- the individual test takers may perform better or worse on a given day
- the markers or the test version may perform differently
- there may be other factors beyond our control.

We aim to produce tests where the overall proportion of score variability caused by ability far outweighs the proportion caused by error.

1.3.2 Reliability in practice

The test developer should be aware of the likely sources of error, and do what is possible to minimise it. Following the procedures and principles described in this Manual will help here. However, using statistics to estimate the reliability of a test's scores is also an important *post-hoc* step. Appendix VII contains more on estimating reliability.

There can be no reliability target for the scores of all tests, because reliability estimates are dependent on how much test taker scores vary. A test for a group of learners who have already passed some selection procedure will typically produce lower reliability estimates than a test on a widely varying population. Also, reliability estimates can depend on the item or task type and the way it is marked. The scores of rated tasks (see Section 5) are typically less reliable than those for DICHOTOMOUS ITEMS because more variance (error) is introduced in the rating process than in the clerical marking process.

However, studying reliability on a routine basis will be useful in identifying particular tests which have worked better or worse, and in monitoring the improved quality of tests over time. Most reliability estimates, such as Cronbach's Alpha or KR-20, are in the range 0 to 1. As a rule of thumb, an estimate in the top third of the range (0.6 to 1) is often considered acceptable.

The statistical estimation of reliability is, however, not usually possible in situations where numbers of test takers, and/or items are low. In these cases, it is not possible to estimate whether reliability is adequate for the purposes of the test. A good assessment strategy in these situations is to consider the test as providing only part of the evidence on which to base a decision. Additional evidence might be gathered from a portfolio of work, multiple tests over a period of time and other sources.

1.4 Ethics and fairness

1.4.1 Social consequences of testing: ethics and fairness

Messick (1989) argued for the critical role of values and test consequences as part of validity. His influence has led to greater attention being paid to the social value of tests as well as their consequences for the STAKEHOLDERS. The effects and consequences of tests include the intended (and hopefully positive) outcomes of assessment, as well as the unanticipated and sometimes negative side-effects which tests might have. For example, the introduction of a new test may affect (positively or negatively) the way in which teachers teach ('washback').

Test providers can conduct washback and impact-related research to learn more about the social consequences of their test. Such research can be done on a very small scale. In a classroom context, it is possible to see whether students start to prioritise certain aspects of the syllabus at the expense of others, perhaps because of the focus of the test. Other ways may be needed to encourage work on the neglected aspects, including shifting the focus of the test.

1.4.2 Fairness

An objective for all test and examination providers is to make their test as fair as possible. See the *Code of Fair Testing Practices in Education* (JCTP 1988) and the *Standards for Educational and Psychological Testing (*AERA et al 1999).

The 1999 *Standards* acknowledge three aspects of fairness: *fairness as lack of bias*, *fairness as equitable treatment in the testing process*, and *fairness as equality in outcomes of testing*.

Kunnan's *Test Fairness Framework* (Kunnan 2000a, 2000b, 2004, 2008) focuses on five aspects of language assessment which need to be addressed to achieve fairness: *validity* (see Section 1.2), *absence of bias* (see Appendix VII), *access*, *administration* (see Section 4) and *social consequences*.

Various bodies have produced Codes of Practice or Codes of Fairness to assist test providers in the practical aspects of ensuring tests are fair.

Test providers can try to minimise bias when designing tests. For example, certain topics (e.g. local customs) may advantage or disadvantage certain groups of test takers (e.g. those from countries with quite different customs). A list of topics to avoid in test items can be given to item writers. Significant groups of test takers may include those defined by age, gender or nationality, though this depends on the testing context (see 3.4.1).

1.4.3 Ethical concerns

Since the early 1980s, ethical concerns have also been discussed in language testing. In particular Spolsky (1981) warned of the negative consequences that HIGH-STAKES language tests can have for individuals and argued that tests should be labelled like medicines: 'use with care'. He focused in particular on specific uses of language tests, e.g. in the context of migration, where decisions made about a person on the basis of a test score can have serious and far-reaching consequences.

The International Language Testing Association (ILTA) published its *Code of Ethics* in 2000; this sets out broad guidelines on how test providers should conduct themselves.

Test providers must ensure that the relevant principles are widely disseminated and understood within their organisations. This will help to ensure that the organisation complies with these guidelines. Further measures may also be appropriate for some aspects of test fairness (see Section 4 and Appendix VII).

1.5 Planning the work

The stages of test development and use form a cycle where success in one stage depends on the outcomes of the preceding stage. It is important, therefore, to manage the whole cycle well. The collection of evidence is also important to consider, as it can be used in making important decisions during the process.

1.5.1 Stages of work

Figure 5 illustrates the stages in constructing a new test. It starts with a decision to provide a test, which may be made by the test provider, or someone else, such as a head teacher, an administration office or ministry. The stage of test development comes next, followed by those stages connected with test use. The completion of each stage rests on the completion of many smaller tasks within that stage. Together, the tasks are designed to achieve the aims listed in the boxes on the right of the diagram. A time line shows that these stages follow each other consecutively because the output from one stage is required for another to start. Once the test has been developed, the stages of test use may be repeated many times using the output (test SPECIFICATIONS) from the test development stage. This allows the production of many EQUIVALENT FORMS of the same test.

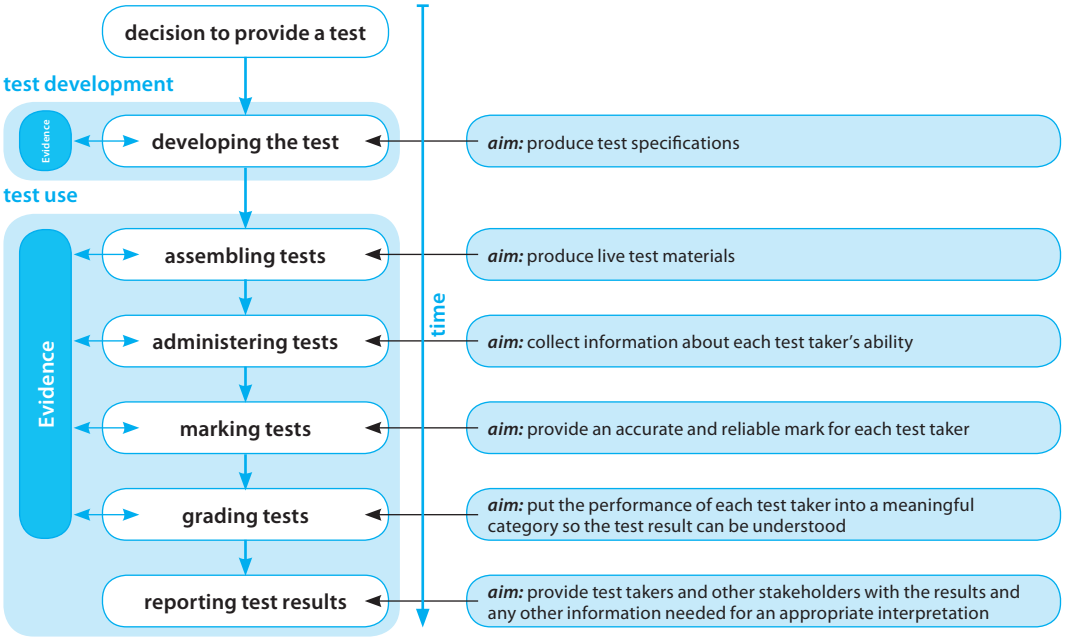


Figure 5 ***The basic testing cycle***

The stages represented in Figure 5 apply to every test construction project, no matter how large or small a test provider may be.

Each of the stages represented in Figure 5 contains many smaller, 'micro' tasks and activities. These are described in more detail in the later sections of this Manual. The process used to complete a task should be standardised to ensure that each time a test form is produced it is likely to be comparable to previous versions.

The collection and use of evidence is also shown in the boxes on the left of the diagram. Evidence, such as background information about the test takers, feedback from people involved, the responses of test takers to tasks and items and the time it takes to complete certain tasks, is important as an ongoing check that the development is proceeding satisfactorily, and also later on in helping to demonstrate that the recommended uses of test results are valid.

Plan to collect and use such evidence as a routine activity. Otherwise it is likely that this important activity will be forgotten during the process of developing the test.

1.6 Key questions

- Which aspects of the CEFR model of language use are most appropriate to your context?
- Which CEFR levels of ability are most appropriate?
- How would you like your test results to be understood and interpreted?
- What may be the greatest threats to reliability in your context?
- How can you help ensure that your work is both ethical and fair to test takers?
- What challenges will you face in planning your testing cycle?

1.7 Further reading

Models of language use
Fulcher and Davidson (2007:36–51) discuss constructs and models further.

Validity
ALTE (2005:19) provides a useful summary of types of validity and the background to the modern conception of validity.

Kane (2004, 2006), Mislevy, Steinberg and Almond (2003) discuss issues relating to validity arguments (which are discussed in Appendix I of this Manual) and give further details concerning how to develop them.

Reliability
Traub and Rowley (1991) and Frisbie (1988) both describe the reliability of test scores in an accessible way. Parkes (2007) illustrates when and how information from a single test can be supplemented with other evidence to make decisions about test takers.

Ethics and fairness
Specialist *Codes of Practice* for language testers have also been developed by professional language testing associations since the early 1990s, for example:

- The *ALTE Code of Practice* (1994)
- The *ILTA Guidelines for Practice* (2007)
- The *EALTA Guidelines for Good Practice in Language Testing and Assessment* (2006).

In the 1990s, a special issue of *Language Testing*, guest-edited by Alan Davies (1997), focused on ethics in language testing and *a language assessment ethics conference* was organised in 2002 in Pasadena. The proceedings of this event led to a special issue of *Language Assessment Quarterly* (also guest-edited by Davies, 2004). McNamara and Roever (2006) provide an overview of fairness reviews and Codes of Ethics for examinations.

Gathering evidence for test fairness and displaying it in the form of an argument is the focus of several articles in *Language Testing*, April 2010 (Davies 2010, Kane 2010, Xi 2010).

2 Developing the test

2.1 The process of developing the test

The aim of developing the test is to produce test SPECIFICATIONS which can be used to construct LIVE TESTS. Test development begins when a person or organisation (the test sponsor) decides that the new test is necessary. Figure 6 illustrates the test development process, showing three essential phases (planning, design and try out) and one phase (informing STAKEHOLDERS) that may be necessary in some contexts. This is because dissemination does not contribute to the production of specifications like the other phases, but aims to inform others about the new test.

A more detailed diagram can be found in Appendix II.

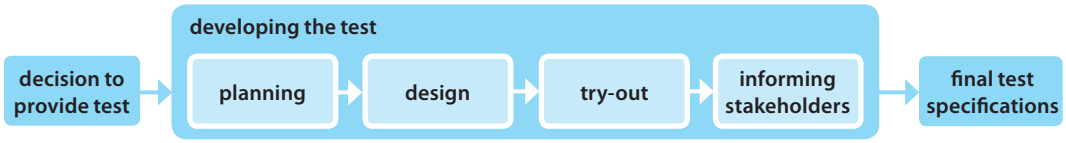


Figure 6 ***The test development process***

2.2 The decision to provide a test

This decision is not considered part of the test development process in this Manual, but it provides important input to the planning phase, as the requirements identified by the test sponsor, who decides that a test is necessary, will determine the test's design and how it is used.

Who decides that a new language test is needed? In some cases, this decision is taken by the test provider who will carry out the test development project. Sometimes the test provider is commissioned to develop the test by a third-party sponsor who has already decided that a test is needed.

In either case, the requirements must be clearly identifiable, and this may involve additional work by those with the task of developing the test. It may be more difficult to understand the sponsor's intentions because they are not in the same organisation, or because they are not language testing or teaching experts and do not know what information the test developer needs.

2.3 Planning

This phase is dedicated to gathering information for use in later stages. Much of this information should come from the sponsor of the test. However, a wider range of stakeholders may also be consulted, including ministries and government bodies, publishers, schools, parents, experts, employers, educational institutions and administrative centres. Where many people are involved structured information may be collected using questionnaires or seminars, whereas in a classroom situation, direct personal knowledge of the context and the test takers may be sufficient. The most important questions that the test developer needs to ask are:

- What are the characteristics of the test takers to be tested? (age, gender, social situation, educational situation, mother tongue, etc.)
- What is the purpose of the test? (school-leaving certificate, admission for a programme of education, minimum professional requirements, formative or diagnostic function, etc.)
- How does the test relate to an educational context? (a curriculum, a methodological approach, learning objectives, etc.)

- What standard is needed for the proposed purpose? (a CEFR level in particular skills, in all skills, a standard relating to a specific domain, etc.)
- How will the test results be used?

The answers to these questions will allow the developer to begin to define the language ability to be tested, to decide how to set cut-off points to make decisions about the test takers (see Section 5) and how to present and explain the results to the people using them (see Section 5).

Questions about the effect of the test on a wider context are also helpful:

- Who are the stakeholders?
- What kind of impact is desired?
- What kind of impact is expected?

Finally, questions of a more practical nature should not be forgotten:

- How many test takers are expected?
- When should the test be ready?
- How will the test be financed and what is the budget?
- How often will the exam be administered?
- Where will the exam be administered?
- What mode of delivery is required? (i.e. paper-based or computer-based)
- Who will be responsible for each stage of test provision? (i.e. materials production and test construction, administration, marking, communication of results)
- What implications will this have for test security? (e.g. should one or more test forms be used?)
- How will the test performance be monitored over the long term?
- Will pretesting be possible or feasible (see Section 3.4)?
- What implications will this have for logistics? (i.e. will the work of the test provider be dependent on other organisations, such as test centres?)

2.4 Design

Information from planning provides the starting point for the design phase. Important decisions about the nature of the test are made and initial test specifications are produced. These specifications describe the overall structure of the test and all aspects of its content. More detailed specifications, for example those required by writers of materials and personnel involved in the delivery and management of the tests, can be developed once the initial specifications have been agreed.

2.4.1 Initial considerations

The first challenge in the design phase is to develop a clearer idea of test content and format. This starts with the information gathered about test requirements and background, such as the characteristics of the test takers, the purpose of the test and the required ability level.

The CEFR is a useful resource for defining the characteristics of the test, with many of its chapters relevant to testing, particularly:

- Chapter 6, on language learning and teaching, invites reflection on learning objectives and teaching methodology, which in turn impact on the style, content and function of tests
- Chapter 7, on TASKS and their role in language teaching, has implications for how tasks might be used in testing
- Chapter 9 on assessment discusses how the CEFR can be used for different testing purposes.

Most obviously relevant are Chapters 4 and 5, which deal with test content and the skills to be tested. These offer the test designer many options to select from within the CEFR's overall action-oriented approach and model of language use (see Section 1.1), relating to, for example:

- the focus of tasks, e.g. showing detailed comprehension of a text, etc. – see CEFR (Ch 4.4 and 4.5)
- what is to be tested, e.g. skills, competences and strategies – see CEFR (Ch 5)
- text types used as INPUT – see CEFR (Ch 4.6)
- text sources – see CEFR (Ch 4.1 and 4.6)
- some indication of topic areas considered suitable for use – see CEFR (Ch 4.1 and 4.2)
- types of PROMPTS used in tests of oral production – see CEFR (Ch 4.3 and 4.4)
- types of real-life situations relevant to the test takers – see CEFR (Ch 4.1 and 4.3)
- the level of performance necessary in those situations – see the numerous illustrative 'Can Do' SCALES in the CEFR
- criteria for assessing free writing tasks and tests of oral production – see the relevant illustrative 'Can Do' scales in the CEFR, e.g. p. 58 and p. 74, etc.

The test provider must also determine a number of technical features of the test, including the following:

- The test's duration – enough time should normally be allowed for an average test taker to complete all the items without rushing. The most important consideration is that test takers have sufficient opportunity to show their true ability. At first this may need to be estimated by an experienced language tester but some exemplars (see Further reading, Section 2.8) may be consulted. After the test is TRIALLED, or used in a live context, the timing may be revised. Sometimes speeded tests are used, whereby a short time limit is set and test takers are encouraged to complete items quickly. In this case, the time limit must also be trialled.
- The number of items in the test – enough items are needed to cover the necessary content and provide reliable information about the test taker's ability. However, there are practical limits on test length.
- The number of items per section – if the test aims to measure reliably different aspects of ability, then a sufficient number of items per section will be needed. Exemplars can be consulted and the reliability calculated (see Appendix VII).
- The item types. Items can elicit selected or constructed RESPONSES. Selected response types include multiple choice, matching or ordering. Constructed response types include short responses (gap-filling exercises) or extended writing. Item types have different advantages and disadvantages. See ALTE (2005:111–34) for more information on item types.
- The total and individual length of texts, e.g. as measured in words. Exemplars (see Further reading, Section 2.8) can provide guidance about how long is practical.
- The format of the test. A DISCRETE-ITEM test consists of short items which are unrelated to each other. In a task-based test items are grouped into a smaller number of tasks, e.g. relating to a reading or listening text. Because task-based tests can use longer, more authentic stimuli they are generally more appropriate for communicative language testing. See ALTE (2005:135–47) for more information on task types.
- The number of marks to give each item and the total marks for each task or COMPONENT – the more marks for each item or section, the greater its relative importance. Generally the best policy is one mark per item. However, there may occasionally be a reason to WEIGHT items by giving more or less than this (see Appendix VII).
- The characteristics of RATING SCALES – will task-specific scales be used, how long will each scale be, will scales be analytic or holistic? (Sections 2.5 and 5.1.3 discuss rating scales in more detail.)

The design phase thus ends in initial decisions concerning the purposes the test is to serve, the skills and content areas to be tested, and details of the technical implementation. We should also consider how the tasks are to be marked, how rating scales for productive skills (e.g. writing and speaking) should be developed (see Section 2.5), how tests should be administered (Section 4) and how markers and RATERS should be trained and managed (Section 5.1.3). All of the stakeholders should then review these proposals in detail, enabling proper evaluation of them.

Communication with test takers and other stakeholders should be considered:

- if formal study is expected, the estimated number of hours of study necessary as preparation for the test
- how specimen papers will be made available
- the information to be given to users of the test (all relevant stakeholders) both before and after the test.

Finally, consider stakeholders' expectations:

- how will the test fit into the current system in terms of curriculum objectives and classroom practice?
- what will stakeholders expect the test to be like?

Chapter 4 of the CEFR provides a particularly useful reference scheme against which the distinctive features of any test in the process of development can be brought into clearer focus. In order to do this, a summary of the test is compiled in diagram form. This approach is exemplified in Appendix III of this Manual. The example test is aimed at B2 test takers studying the language in a business context, and is made up of four components (or 'papers'). There is both an overview of the content of the whole examination, and a general description for a single paper within the examination.

2.4.2 How to balance test requirements with practical considerations

At this stage of test development the proposed test design must be balanced against practical constraints. Information about constraints is gathered in the planning phase, at the same time as test requirements (Section 2.3). The test developer must reconcile requirements and constraints, and this must be agreed by the test sponsor. Bachman and Palmer (1996:Ch2) provide a framework to do this through their concept of test usefulness. In their view, test usefulness is a function of six qualities:

- Validity – the interpretations of test scores or other outcomes are meaningful and appropriate.
- Reliability – the test results produced are consistent and stable.
- Authenticity – the tasks resemble real-life language events in the domain(s) of interest.
- Interactivity – the tasks engage mental processes and strategies which would accompany real-life tasks.
- Impact – the effect, hopefully positive, which the test has on individuals, on classroom practice and in wider society.
- Practicality – it should be possible to develop, produce and administer the test as planned with the resources available.

These qualities tend to compete with each other: for example, increasing task authenticity may lead to a decrease in reliability. For this reason, it is the effort to find the best balance between them which is important in increasing test usefulness overall.

2.4.3 Test specifications

The output from the test development is a set of finished test specifications. The first draft of these specifications contains decisions about much of the information already discussed. After the Try Out (see Section 2.5), these specifications will be finalised. If the test is high-stakes, specifications are particularly important because they are an important tool to ensure the quality of the test and to show others that the recommended interpretations of the test results are valid.

Specifications are also important for low-stakes tests, as they help to ensure that test forms have the same basis and that the test correctly relates to a teaching syllabus, or other features of the testing context.

Test specifications may be written in different ways according to the needs of the test provider and the intended audience. A number of models of test specifications have been developed (see Further reading, Section 2.8) and can serve as a useful starting point.

2.5 Try-out

The aim of this phase is to 'road test' the draft specifications and make improvements based on practical experience and suggestions from stakeholders.

Once the specifications have been drafted, sample materials are produced. This can be done following guidance in Section 3 of this Manual. Information about these materials can be gathered in a number of ways:

- PILOTING (asking some test takers to sit the test) and the analysis of the responses (see Section 3.4 and Appendix VII)
- consulting colleagues
- consulting other stakeholders.

Piloting should be conducted using test takers who are the same or similar to the target test takers. The pilot test should be administered under exam conditions, as the live test would be. However, piloting will still be useful even if it is not possible to replicate a live test administration (perhaps there is insufficient time to administer the entire test, or for some other reason), or if numbers of test takers are small. It can provide information about the timing allowance needed for individual tasks, the clarity of task instructions, appropriate layout for the response, etc. For oral components, observation of the performance (e.g. via a recording) is recommended.

Consultation with colleagues or stakeholders can take a variety of forms. For a small group, face-to-face communication is possible. However, for larger projects, questionnaires or feedback reports may be used.

Information from piloting also allows fairly comprehensive MARK SCHEMES and rating scales to be devised (see Section 5.1.3 for features of rating scales). Features can be identified in test taker performances which exemplify the ability levels best. They can form the basis of descriptors at each level. Once constructed, rating scales should be piloted and the way in which they are used by raters analysed, qualitatively or quantitatively (see Appendix VII). Further piloting and revision may be necessary.

Other kinds of research may be needed to resolve questions which have arisen during the try-out phase. It may be possible to use piloting data to answer them, or separate studies may be needed. For example:

- Will the task types we wish to use work well with the specific population the test is aimed at (e.g. children)?
- Will task types be valid in the domain the test is aimed at, e.g. tourism, or law?
- Do items and tasks really test the skill area they are supposed to? Statistical techniques can be used to determine how well items and tasks test distinct skill areas (see Appendix VII).
- Will raters be able to interpret and use the rating scales and assessment criteria as intended?
- Where a test is being revised, is a comparability study needed to ensure that the new test design will work in a similar way to the existing test?
- Do items and tasks engage the test taker's mental processes as intended? This might be studied using verbal protocols, where learners verbalise their thought processes while responding to the tasks.

Specifications may undergo several revisions before they reach the form they are to take for the live test.

2.6 Informing stakeholders

Test specifications have many uses, as they may be read for purposes such as writing items, preparing for the test and deciding what to teach. In many contexts, this will mean that different versions should be developed for different audiences. For example, a simplified version listing the linguistic coverage, topics, format, etc., of the test could be produced for those preparing for the test. A far more detailed document may be used by those writing items for the test.

In addition to test specifications, stakeholders will find it very useful to see sample materials (see Section 3 for more information about producing materials). If it is appropriate, the sample materials can include not only question papers, but audio or video recordings used in listening tasks. In a classroom context, these materials could be used as part of the preparation for the test. In future years, used test versions can be used as sample materials.

For speaking and writing tasks, it may be useful for test takers to see responses to the sample materials. These can be prepared if the materials have been trialled, or used as a live test in the past. Alternatively, advice to test takers can be provided to help them prepare for the test.

Any materials must be available in advance of the time they are needed. If any other material is needed, such as regulations, roles and responsibilities, and timetables, it should also be prepared in advance.

2.7 Key questions

- Who decided that there should be a test – what can they tell you about its purpose and use?
- What will the educational and social impact of the test be?
- What type and level of language performance needs to be assessed?
- What type of test tasks are necessary to achieve this?
- What practical resources are available? (e.g. premises, personnel, etc.)
- Who should be involved in drafting test SPECIFICATIONS and developing sample test materials? (e.g. in terms of expertise, influence, authority, etc.)
- How will the content, technical and procedural details of the test be described in the specifications?
- What sort of information about the test needs to be given to users? (e.g. a publicly available version of the test specifications, and how this should be distributed)
- How can the test be tried out?
- How can stakeholders be best informed about the test?

2.8 Further reading

There are a large number of CEFR-related samples of test material to assist those involved in test development in understanding the CEFR levels. See Council of Europe (2006a, b; 2005), Eurocentres/Federation of Migros Cooperatives (2004), University of Cambridge ESOL Examinations (2004), CIEP/Eurocentres (2005), Bolton, Glaboniat, Lorenz, Perlmann-Balme and Steiner (2008), Grego Bolli (2008), Council of Europe and CIEP (2009), CIEP (2009).

Sample test specification formats may be found in Bachman and Palmer (1996:335–34), Alderson, Clapham and Wall (1995:14–17) and Davidson and Lynch (2002:20–32).

A number of grids are available which provide templates for describing and comparing tasks, including ALTE Members (2005a, b; 2007a, b), Figueras, Kuijper, Tardieu, Nold and Takala (2005).

3 Assembling tests

3.1 The process of assembling tests

The aim of the assembling tests stage is to provide live test materials, made according to the test SPECIFICATIONS and ready on time. The process of assembling tests breaks down into three broad stages, illustrated in Figure 7.

Figure 7 ***Broad stages in the test assembly process***

Producing test papers and constructing tests are described here as two distinct stages, as it is easier to make the aims of each step clear. However, if test materials are produced and formed into a test paper at the same time, the principles are the same. A quality control stage should follow, with the possibility to change the test materials where necessary.

3.2 Preliminary steps

Before starting materials production the following preliminary steps must be considered:

- item writer recruitment and training
- materials management.

3.2.1 Item writer recruitment and training

Item writers may also be the same as those developing the test. In this case, recruitment is not necessary and training is relatively easy because they are already familiar with the test and its aims.

Where item writers need to be found, the test provider must decide what minimum professional requirements item writers need to fulfil. These can include level of language ability and understanding of the testing context. Other important aspects for item writing, like knowledge of an existing test, or of principles of assessment, can be dealt with in training (see ALTE 2005), so do not need to be included in the minimum requirements. Ongoing training, monitoring and evaluation will help to ensure the continuing professional development of the item writer.

Language teachers often make good item writers as they have developed a deep understanding of language learners and of the language. They will be even more suitable if they have prepared students to take a similar test, or are involved in marking or oral examining. Item writers may write for all parts of the test, or just for specific components, depending on their knowledge and the requirements of the test provider.

3.2.2 Managing materials

The test provider must set up a system to collect, store and process the items. This is particularly important when a large number of items and tasks are involved. All test materials must go through the same process of quality assurance, such as editing and piloting. For this reason, it should be possible to see at what stage any item is in the process at any time. This becomes increasingly important as larger numbers of items are produced and more people are involved at different stages. A basic system of materials management should include:

- a unique identification number for each item
- a checklist which records the stages completed, changes and other information
- a way to ensure that items and related information are accessible and that versions from earlier editing stages are not in circulation – perhaps by storing them in a single, central location, or by formally passing everything on by email at the end of each stage of writing and editing.

3.3 Producing materials

Item writers are asked to produce materials which will be used in the live test. In this Manual, this step is referred to as 'commissioning'. Appendix IV contains information which may help them in this task. They must know how many items are needed, of what types, and when to have them ready.

This section focuses on determining what materials are required and communicating this to item writers. Deciding when they should be ready is done by working back from the date of the live test.

3.3.1 Assessing requirements

In order to construct a test, test providers must have a choice of tasks and items to select from. It is difficult to know precisely how many tasks to commission, as the constructed test needs to balance a range of features: item type, topic, linguistic focus and difficulty (see Section 3.5). This means that more items must be commissioned than will actually be used. Another reason for commissioning excess items is that some items are almost certain to be rejected at the quality control stage.

3.3.2 Commissioning

Materials may be commissioned for a particular test administration, or to add new material to an ITEM BANK, from which tests will later be constructed. In either case, sufficient time for all steps in test production must be allowed.

A number of parameters for producing items should be agreed and recorded. The key aim is to avoid confusion and misunderstanding. A longer list of formal requirements and requests is useful if the group of item writers is large and diverse. However, agreeing and recording a small number of the points below will be useful in any situation:

Details of the materials required
This will include details:

- of the number of texts, tasks and items required
- for texts: whether to write items immediately, or only after acceptance of the text
- for speaking tests with visual prompts: whether a visual prompt is expected, or an indication of what sorts of visual prompts are needed
- of issues regarding copyright of pictures or texts and how to deal with them

and requests:

- for a KEY or mark scheme for each item, including the correct response
- for writing tests: sample answers to ensure that the task can be completed within the word limit and at the language ability level of the potential test takers
- for a completed form which describes the task in a standard way.

Details of how materials should be presented

- Electronic copy is the most useful format: it can be easily stored and the item writer can enter it into a template, which will ensure consistency of formatting.
- If writing a whole paper, consider whether items should be numbered consecutively throughout and the sections run on after each other, or whether each section or exercise should be on separate sheets.
- Consider the way in which submissions should be labelled, e.g. with the name of the item writer, date and test name.

(All the details above can be covered in the guidelines for item writers – see below.)

Details of the deadline by which materials must be submitted
Item writers should be told when editing of their material will take place and whether they are expected to take part in editing. If item writers are not involved in the remainder of the test production process, they can also be told how their role fits into the overall production schedule, as this will help them to understand the importance of keeping to agreed deadlines.

Other details, such as terms of employment
Some item writers will need to know the terms of their employment if item writing is something additional to other work for the organisation, or if freelance item writers are used. A fee may be payable for accepted materials only (with no payment for rejected materials); alternatively, a fee may be paid on initial submission of the materials, with a further fee for all materials accepted. Rates for various types of item may vary, or there may be a sum paid for a complete section or test.

Teachers who have been asked to write materials for school tests will need enough time within the school timetable to develop the materials.

The following documents should be given to item writers:

- Detailed test specifications suitable for item writers. These may be confidential documents, going into more detail than is made available publicly, and should contain detailed advice on the selection and presentation of materials. This can prevent item writers from wasting time by making their own, possibly mistaken, assumptions about what is acceptable.
- Sample materials or past papers.

Item writers should also be told the profile of the target test takers, including factors such as age, gender and linguistic background.

Depending on the testing context there may be a need for additional documents and guidance, such as:

- a form for the item writer to sign to indicate acceptance of the commission
- an agreement which indicates that the test provider will own the copyright of the test materials
- a list or lexicon defining the range and level of vocabulary and/or structures to be used
- a handbook giving information about the test provider.

3.4 Quality control

3.4.1 Editing new materials

Once test materials have been submitted, they must be checked for quality. This is done using expert judgement and by trying them out. When one or more changes are made to an item or task, it should be checked again to see if it is suitable.

This kind of checking is essential but each item or task should, ideally, not be checked by its author. In contexts where resources are short, items and tasks may be checked by a small group of colleagues. However, if the item writer is working alone, leaving extra time between producing the tasks and reviewing them, and reviewing many items at once can help them to be more objective.

The very first check should be that the materials conform to the test specifications and any other requirements

set out when commissioned. Feedback should be given to item writers to allow them to revise their work and to develop their skills as item writers. This can also include suggestions on how the item might be changed (see Appendix V).

Texts may be commissioned *without* items. If the text is accepted the item writers then produce items for the second stage of editing.

This initial check can be quite quick and many items can be reviewed in a relatively short space of time. If there are many items to check, this stage could take the form of a separate meeting.

The next stage of editing involves a more detailed review of each item or task. It is important that each is reviewed by someone other than its author. Within a school, for example, teachers writing tests for their own classrooms could help to review each others' items.

Having more than four or five people in an editing group tends to make the process slow, while fewer than three may not bring enough variety in points of view. If several meetings are required, one person within the organisation may be designated as co-ordinator. This person should schedule meetings, decide who attends each and which materials are discussed in each meeting.

Participants can review materials in advance to save time at the meeting by working through them in the following way:

- TEXT-BASED ITEMS should be read *before* reading the text, as this helps to highlight items which can be answered *without* reference to the text (e.g. solely on the basis of common sense or background knowledge).
- All other items can be answered, without looking at the key, as if taking the test. This will help to identify items where there is more than one possible correct answer, unclear or badly phrased options, an implausible distractor, or items which are difficult or unclear.
- Reading and listening texts should be checked for their length, suitability of topic, style and level of language. Checking the level of language requires expert judgement, and may be assisted by referring to linguistic descriptions.

If editing in a group, any problems observed in the materials can be raised and discussed in detail within the group. There is often a lot of discussion about materials, and item writers need to be able to accept as well as offer constructive criticism, which can be difficult to do. If an item writer finds it necessary to justify and explain a piece of material to experienced colleagues, then it is likely that the material is flawed in some way.

One person in the group should take responsibility for keeping a detailed and accurate record of all decisions made about materials, showing clearly any changes made at editing. At the end of the meeting, it is vital that there should be no doubt about what changes were agreed.

The test provider should make final decisions and decide when there has been enough discussion.

The following points apply to editing meetings focusing on the detail:

- Special attention should be given to RUBRICS (the instructions given to test takers with items) and keys.
- Items which may introduce bias into the test can be identified by referring to a list of topics to avoid or other aspects to be aware of (see Appendix VII).
- Some materials may appear to have potential, but need more changes than is possible during the meeting. These are given back to their original writers for further work or may be given to a more experienced writer for revision and further editing.
- After the meeting, spare and used copies of the edited materials should be destroyed for security reasons. The amended copies of accepted materials are kept by the test provider.
- Item writers should get feedback from the test provider on rejected material, especially if they have not taken part in editing, or if they have not been present during editing of their own materials.
- Editing meetings are an excellent opportunity for new item writers to learn from working in a group with more experienced writers.

3.4.2 Piloting, pretesting and trialling

Some form of try-out is needed as test takers can still respond to items in unexpected ways. Piloting, PRETESTING, TRIALLING, or a combination of these methods, are used, depending on the aims and resources of the test provider.

Piloting involves asking a small number of people to complete items as if in a test. This may be quite informal and could, for example, involve work colleagues if no one else can be found. Their RESPONSES are analysed and, together with their comments (see Appendix VI), may be used to improve items further.

Pretesting is usually conducted for tests which use OBJECTIVELY MARKED items. Under pretesting, live test conditions are observed, test takers who are similar to the expected test taker population are chosen, and sufficient numbers of responses are collected to enable statistical analysis (see Appendix VII). The analysis can reveal how well options worked, how hard an item was, what the average score was, whether the test was at the right level for the group of test takers who took it, how much error can be found, whether items may be biased (see Appendix VII), whether they contribute to measuring the same construct, their relative difficulty, and other information. The most basic types of statistical analysis (see Appendix VII) can be extremely informative and can be attempted using inexpensive software that is not difficult to use.

Responses to subjectively marked tasks (those testing writing and speaking) can be analysed statistically but a qualitative analysis based on fewer responses may be more informative. Small-scale pretesting of productive tasks is sometimes called *trialling* to distinguish it from the pretesting of objectively marked tests. It can show whether the test tasks work satisfactorily and elicit the kind of performance expected.

Unlike piloting, a pretesting programme is like a live test programme, as it requires similar resources. These include:

- sufficient numbers of test takers (see Appendix VII)
- securely printed test papers
- test venues and staff
- markers.

Pretest test takers should be as similar as possible to test takers who will actually take the live test. A good solution, if practical, is to recruit individual learners who are preparing to take a live exam.

Offering feedback on performance is a good way to motivate test takers to participate and to provide responses which represent their true ability. Feedback can help test takers and their teachers to understand their current level of proficiency and to target areas requiring remedial work before taking the live test.

A potential disadvantage in using such test takers is that exposing items may threaten the security of future live tests. For some test providers, this problem can make any kind of pretesting difficult to carry out.

To reduce the risk complete test papers should not be pretested in the same form that they will appear in live tests. TEST CONSTRUCTION schedules should also attempt to build in a significant time lag between the pretesting of an item and its use in a live test. Where pretests are administered externally, staff involved should be given instructions on how to keep the pretests secure and asked to sign confidentiality agreements.

A pretest paper need not closely resemble the actual live test paper, as the items, not the test itself, are being pretested. However, if pretest test takers are motivated by the chance to experience taking a test which is similar to the live one (i.e. as a way of preparing for it), using a format which is very similar to the live test is advisable.

In all cases, the pretest should be administered under live test conditions. If test takers are distracted, cheat or are allowed different lengths of time, it may result in data which is difficult to interpret.

When the quality of statistical information is particularly important (e.g. the items are being CALIBRATED – see Appendix VII), the numbers of test takers should be sufficient for the purpose. The minimum number to aim for will depend on the types of analysis to be employed. But even a small sample of test takers (below 50), may provide useful information to indicate problems with some items. With smaller samples, qualitative analysis will be more useful.

It is also more crucial to find test takers who are as similar as possible to live test takers. Smaller, less representative samples of test takers will allow more tentative conclusions to be drawn from the analysis, which must be balanced with expert judgement during a review of items. See Appendix VII for more on analysis.

If pretesting is done to gain qualitative information about items, some consideration is required in order to gather the maximum amount of useful information:

- For objectively marked items, it may also be possible to gather feedback from test takers and teachers. In order to facilitate this, a list of questions or a questionnaire may be of use (see Appendix VI).
- For tasks in speaking tests where an interlocutor is involved, his or her feedback may also be useful to collect. It can show the test provider whether the task was understood by the students, whether it was suitable for their experience and age group and whether they were provided with enough information to fulfil the task adequately (see Appendix VI).
- For subjectively marked items and tasks, test taker responses can show whether test takers were given the opportunity to show the range of discourse structure, syntactic structure and vocabulary expected at this level.
- Test taker feedback on the pretesting experience as a whole may also be gathered (see Appendix VI), as may other feedback related to the entire session.

3.4.3 Review of items

A review meeting should follow a piloting or pretesting session. Attendees at such a meeting may include the test provider, experienced item writers and, in the case of subjectively marked items and tasks (e.g. writing tests), an experienced rater.

The aim of this meeting is to use the evidence from piloting or pretesting to retain, improve or reject items. This is illustrated in Figure 8, where items sent for improvement are piloted or pretested a second time.

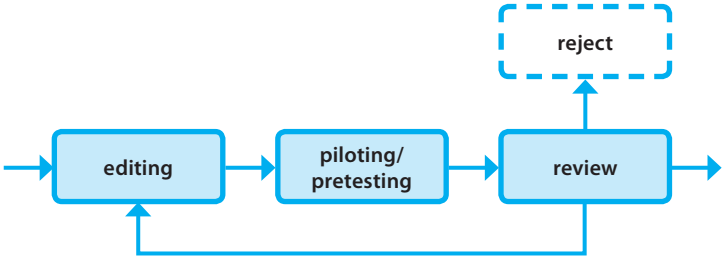


Figure 8 ***Improving items with quality assurance***

Pretest review addresses the following general points:

- which material is ready to go into a live test?
- which material should be rejected as unsuitable?
- which material could be rewritten and pretested again before being reconsidered for inclusion in a live test?

The review meeting should consider the following:

- How well the test takers matched the target population for the live test. This may give a sense of how far the evidence from the analysis can be trusted.
- How engaging and accessible the tasks and topics were and whether there were any problems with any of the administrative procedures.
- The performance of individual items and tasks. When evaluating subjectively marked tasks, it is useful to have a selection of test taker responses available for review. For objectively marked items the statistical analysis may reveal problems with items which expert review can confirm and correct. However, where the analysis is based on poor data (e.g. with unsuitable test takers, or small numbers) it should be treated with caution. Other information, such as qualitative appraisal of the items and tasks, should be given more importance.

- Where statistical and other information is to be held against tasks in an item bank, a coherent and consistent approach should be followed. This will help to ensure its usefulness at the test construction stage. See Appendix VII for more information about statistical analysis.

3.5 Constructing tests

Once sufficient materials are available, tests can be constructed. The aim of this phase is to produce test forms to the desired quality standards and according to the requirements of the test specifications.

The TEST CONSTRUCTION stage involves balancing a number of different aspects, such as test content and item difficulty, so that the test as a whole meets the required specification.

Certain features of a test may be fixed on the basis of the test specifications and format (e.g. the number and type of items/tasks to be included), while other features may remain relatively flexible within defined limits (e.g. topics, variation in accents, etc.). Guidelines can help to achieve an appropriate balance between the following features:

- level of difficulty (this may involve a subjective judgement, or where an item banking approach is followed, can be described by the MEAN difficulty of the test items and the RANGE of difficulty covered – see Appendix VII)
- content (topics or subject matter)
- coverage (representativeness of tasks in relation to the construct)
- gradedness (whether the test becomes progressively more difficult).

These guidelines should be applied to the entire test, looking across and comparing the individual components.

Additional considerations may apply for certain test types. For example, in a reading test containing several texts and items, a check should be made to avoid duplication of text topics or excessive length in terms of total number of words overall. Similarly, in a listening test, it is important to check that there is a suitable balance of male/female voices and regional accents (if relevant).

3.6 Key questions

- How will the test materials production process be organised?
- Can some form of item bank be used?
- Who will write the materials?
- What should the professional requirements for item writers be?
- What training will be given?
- Who will be involved in editing meetings?
- How will editing meetings be managed?
- Is it possible to PRETEST or TRIAL test materials?
- What might be the consequences of not pretesting or trialling and how might these be addressed?
- What type of analysis is to be done on the performance data gathered through pretesting?
- How will the analysis be used? (e.g. for TEST CONSTRUCTION purposes, for test writer training, etc.)
- Who will be involved in the activity of test construction?
- Which variables need to be considered and balanced against each other? (e.g. level of difficulty, topical content, range of item types, etc.)
- What will be the role of statistical analysis? (e.g. in establishing the mean difficulty and difficulty RANGE of the test)

- When making decisions, how important will statistical analysis be in relation to other information?
- Will the constructed test be VETTED independently?
- How will the constructed test form be matched to other forms of the same test or made to fit within a larger series of tests?

3.7 Further reading

For item writer guidelines, see ALTE (2005).

For the analysis of test tasks, see ALTE (2004a, b, c, d, e, f, g, h, i, j, k).

Linguistic descriptions of some languages that relate to the CEFR are available: Reference Level Descriptors (RLDs) (Beacco and Porquier 2007, 2008; Beacco, Bouquet and Porquier 2004; Glaboniat, Müller, Rusch, Schmitz and Wertenschlag 2005; Instituto Cervantes 2007; Spinelli and Parizzi 2010; **www.englishprofile.org**). Threshold (van Ek and Trim 1991), Waystage (van Ek and Trim 1990) and Vantage (van Ek and Trim 2001) are forerunners of RLDs.

Appendix VII contains more information about the way in which statistical information can be used when assembling a test.

4 Delivering tests

4.1 Aims of delivering tests

The main aim of the test delivery process is to collect accurate and reliable information about the ability of each test taker.

The big challenges in test delivery are logistical and not related to improving the quality of test materials as in previous stages. Test providers must ensure that:

- a test taker's performance on the test is influenced as much as possible by their language ability and as little as possible by irrelevant factors, such as a noisy environment, or cheating
- a test taker's RESPONSES and marks are collected efficiently, securely and are available for the next stages of marking and GRADING
- all test materials arrive in the right place at the right time.

These tasks are important whether the test is on a large or small scale. Something as simple as checking the suitability of a room before the test can make an important difference.

An additional aim could be to collect more information about the background of test takers. This is particularly important where test takers are not already known by the test provider. This information can provide a better understanding of who takes the test and can contribute evidence of validity (see Appendix I and Appendix VII).

4.2 The process of delivering tests

The test delivery process is represented in Figure 9. Several of the stages, such as test taker registration or supply of materials may seem very straightforward in some contexts, such as the administration of a classroom test. However, care should still be taken that venues are suited to testing and external influences, such as noise, are reduced. In other contexts, however, the logistics are considerably more challenging and these stages will need greater effort to complete successfully.

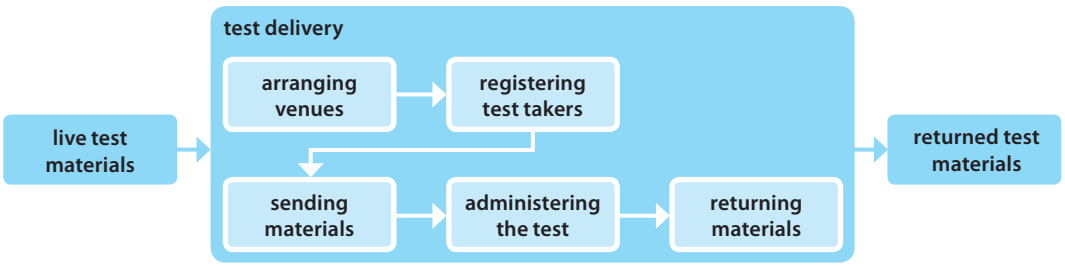


Figure 9 ***The test delivery process***

4.2.1 Arranging venues

The venue(s) where examinations are administered should be inspected in advance. This may be done directly by the exam provider, or by a third party, such as a trusted person in a school which administers the exam. Where other organisations are asked to administer exams, they should be approved. They may be judged against criteria such as:

- capacity to process the numbers of test takers expected
- access to appropriate venues

- security of storage facilities
- willingness to adopt regulations of the test provider
- willingness to train staff to follow the test provider's procedures.

If some centres are arranged by third parties, the test provider should consider setting up a system of random inspections to check the quality of administration work done on their behalf.

When the venues are inspected, the same criteria should be used each time. It is probably better to check the venues before each administration, as something may have changed without the exams administrators being notified, such as the presence of building works within the vicinity.

The features to be checked include:

- ambient noise
- internal acoustics (especially for listening tests)
- size (ability to accommodate the numbers required with sufficient spacing between desks)
- room shape (in order to allow INVIGILATORS to see all test takers clearly)
- accessibility
- the availability of facilities, such as toilets, or a waiting area for test takers
- secure storage facilities for exam materials before and after the administration.

Venues which are found to be unfit for purpose or organisations which make serious errors in administration may be removed from the list of possible venues or collaborators.

4.2.2 Registering test takers

For classroom testing, a list of students in the class and personal knowledge of them may be sufficient. However, where the test takers are unknown to the test provider, or where additional test takers can enrol, information about the test takers should be collected. A registration process provides the information necessary to administer the tests and to process and deliver the results. Test takers can also request special arrangements at this time if they have a particular disability, such as:

- deafness or poor hearing
- partial-sightedness or blindness
- dyslexia
- reduced mobility.

Requests for special arrangements should be properly assessed and, where appropriate, some form of assistance and/or compensation made. For this reason, it is better to have standard procedures for the most common requests. These procedures should include the kind of evidence the test taker should provide (e.g. a letter from a doctor), the actions that may be taken and the date by which the request should be received.

For some special needs, such as those of a test taker who is unable to walk and needs help to take his or her place in the examination room, assistance may be easy to provide.

Some other measures need to be considered more carefully, however. Alternative papers or special aids could be provided to test takers who have difficulty reading, such as the dyslexic or partially sighted. Differences in administration must not, however, advantage some test takers in comparison to others.

It is also possible to collect background information about the test takers at this stage. Test taker characteristics may be used to draw important conclusions about the comparability of different groups of test takers sitting the test. It may include:

- educational background
- native language

- gender
- age
- target language-learning experience.

In all cases, the reason for the collection of this information must be made clear to test takers and all personal data which is collected must be maintained securely to ensure test takers' rights to privacy are protected.

In addition to collecting information, registration is also a good opportunity to provide information to test takers. They should be told the conditions of registration, the rules when sitting the exam, the possibilities for appeal, how and when to request special assistance, etc. Test takers must be told the venue and time of the exam, as well as any other practical information which is available at the time of registration. In order to ensure that all test takers receive complete and correct information, it can be provided in printed form, on the internet or through a standard email.

Registration may be carried out by the exam provider directly, those administering the test, or independent bodies, such as a ministry of education. As far as possible, the test provider should ensure that registration is uniform for all test takers.

4.2.3 Sending materials

Materials may need to be delivered to exam venues. The method of delivery must be both timely and secure so that all materials are in place by the time of administration.

It is often better to send materials well in advance of the exam date to ensure that there is no danger they will be late and that, if materials are missing, they can be replaced. However, the test provider should be satisfied that materials will be kept securely at the venue for the entire time they are there.

Administrators should check the content of the despatch on receipt against a list of what is required. If something is missing, or there are damaged items, the administrators should follow agreed procedures and request replacements or additional materials.

4.2.4 Administering the test

A sufficient number of invigilators, raters and other support staff must be arranged for the day of the examination. Everyone who is involved in administering the exam should understand what their responsibilities are beforehand. Where many staff, rooms or sessions are involved, information can be distributed in the form of a timetable.

The test administration guidelines should include instructions for checking of test takers' identity documents and whether to admit latecomers.

Before the beginning of the exam, clear instructions should be given to test takers about how to behave during the examination. This may include information about unauthorised materials, the use of mobile phones, leaving the room during the exam, and the start and end times. Warnings against unauthorised behaviour such as talking and copying should also be given.

During the examination, invigilators should know what to do if regulations are broken, or there are other events, foreseen or unforeseen, e.g. if test takers are found cheating, if there is a power failure or if something else happens that may cause bias or unfairness or force the session to stop. In the case of cheating, invigilators should be aware of the possible dangers from devices such as digital sound recorders, MP3 players, scanning pens, and mobile telephones with cameras.

In the case of unforeseen events, the invigilator could be asked to use their judgement and to submit a full report to the exam provider. This report should include detail such as how many test takers were affected, the time the problem was noticed and a description of the incident. A telephone number may also be provided which invigilators could ring for advice in the event of an emergency.

4.2.5 Returning materials

Examination papers should be packed and returned to the testing body, or destroyed. If returned, they can be sent together with any relevant documentation, such as attendance registers and room plans as soon as the session has finished. Materials should be returned by secure means, probably using the same service that delivered them. This service would preferably provide a tracking facility in case materials are delayed or go missing.

4.3 Key questions

- What resources are available to administer the test? (administrative staff, invigilators, rooms, CD players, etc.)
- How should staff be trained?
- How can resources such as rooms and CD players be checked before the day of the exam?
- How frequent are sessions?
- How many test takers are expected?
- How will test takers be registered and their attendance at the exam be recorded?
- How many venues are being used? If there is more than one venue, are they far apart or difficult to reach?
- How will materials be transferred to and from the venue(s)?
- How will test materials be securely stored before administration?
- What could go wrong? Are there procedures and regulations about what to do if this happens?

4.4 Further reading

See ALTE (2006b) for a self-assessment checklist for logistics and administration.

5 Marking, grading and reporting of results

The aim of marking is to assess every test taker's performance and provide an accurate and reliable mark for each. Grading aims to put each test taker into a meaningful category so that their test score can be more easily understood. A meaningful category might be one of the Common Reference Levels of the CEFR, such as A2 or C1. When reporting results, the aim is to provide the test taker and other STAKEHOLDERS with the test result and any other information they need to use the result in an appropriate way. This may be to make a decision about the test taker, such as whether to offer them a job. Figure 10 gives an overview of a common version of the process. In some cases, however, the evaluation of test taker performance might be done at the same time as the exam. For example, speaking ability is sometimes assessed in this way, although marks may be adjusted by the test provider before the reporting of results.

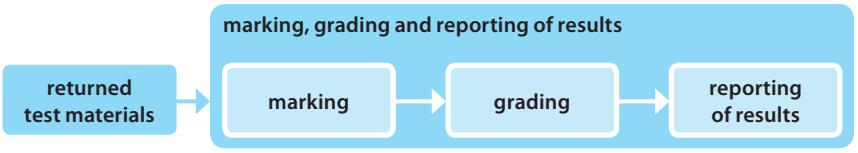


Figure 10 ***The process of marking, grading and reporting of results***

Preliminary steps

Before undertaking marking and grading the following steps must be taken:

- develop the approach to marking
- recruit MARKERS and RATERS
- train markers and raters.

5.1 Marking

While the expression *marking* covers all activities by which marks are assigned to test RESPONSES, a distinction is often made between the *marker*, indicating a less skilled role, and the *rater*, which is a role requiring professional training. This is the distinction we observe in this text. This section covers clerical (i.e. human) and machine marking.

5.1.1 Clerical marking

CLERICAL MARKERS do not need to be testing experts – having a high level of proficiency in the language tested is a sufficient qualification. However, clerical markers need training, guidance and unambiguous answer keys to do a good job. If marking is done by a small group of colleagues, they can check the quality of each others' work.

The process of marking must be managed so that procedures are carried out according to plan and results are ready when required, but also so that the workload for each marker is never so high as to threaten reliability or accuracy.

Marker recruitment and training

Clerical marking at its simplest involves matching the test taker's response to one or more set answers. Item types such as multiple choice give the clearest example, where no deviation from options given is permitted. Where this type of marking is required, the marker must be literate in the language concerned, attentive to detail and prepared to engage in repetitive activity, but they need no other special skills. Training for such a marker would involve familiarisation with the procedures to follow. Given appropriate technology this kind of marking can be done equally well, or better, by machine.

Where items require more than matching a response with a set answer, the marker may need to have some knowledge of the language, language learners and the construct of the test. For example, PARTIAL CREDIT ITEMS carry a range of marks depending on degree of success. For example, one mark may be awarded for selecting a suitable verb and a further mark for using the correct form. Some level of relevant expertise is needed for markers to distinguish between a correct and an incorrect response.

For such items it may also be difficult to ensure that the key is exhaustive. Therefore it is helpful if the marker can identify and report alternative responses they find.

Where markers are employed on a temporary basis but may be re-recruited for future sessions, an evaluation system based on a number of parameters, such as accuracy, reliability and speed, is useful. Markers found to be unsatisfactory may be rejected or retrained. Such a system can be connected with training, as represented in Figure 11. Markers who return frequently may not need to participate in all training. Appraising their performance (see 5.1.3, Monitoring and quality control) will make it easier to decide whether a marker should be asked to return for another session, return and undergo some additional training, or be replaced.

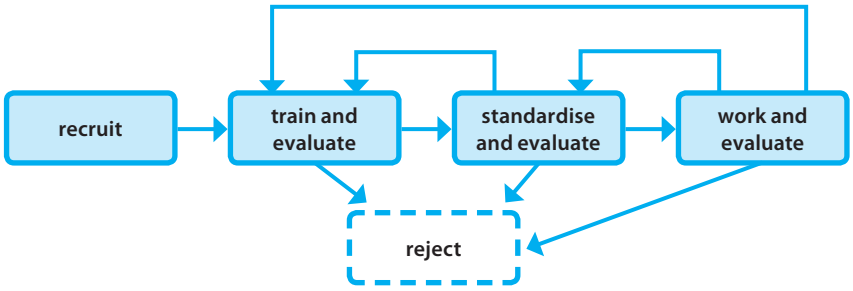


Figure 11 ***Marker/rater recruitment, training and evaluation***

Guidance on evaluating responses

A formalised answer key is the best way to record and to communicate the correct answers to markers. Item keys are developed at the same time as the item and undergo the same editing procedures. The key should cover acceptable responses as comprehensively as possible and be unambiguous for markers to interpret.

Figure 12 shows an example where test takers are asked to fill the gap using the word given ('like'). The key provides four possible alternatives for one element (1 mark) and one possibility for the second (1 mark). The total number of marks available for this item is, therefore, 2.

Clear layout of the answer key and other documents will lead to more efficient, accurate and reliable work from the markers.

The shop will close down whatever our feelings may be.

like

The shop is …… or not.

Key:

(going/sure/certain) to close down/closing down	1
whether we **like** it	1

Figure 12 ***Gap-fill item example***

There may be other responses which are correct but have not been included in the answer key. For this reason, markers should be asked to record all alternative answers they find which they feel are correct. These alternative answers should be evaluated and if they are actually correct, test takers must be credited. If a small group of markers is used, regular discussions with a test constructor may be enough to solve any problems. However, in other situations, if the key is re-evaluated and changed, some or all of the papers will need to be re-marked.

Managing the process of marking

The time to complete marking is normally limited, because results must be issued to test takers by a certain date. The time needed may be estimated by considering the number of test takers and the number of markers available. If possible, it is better to overestimate the time needed slightly, or employ more markers than estimated, to ensure that any difficulties can be coped with.

If large numbers of test takers and markers are involved, a system is needed to track test papers through the process. A simple system might involve recording the number of the SCRIPT together with the number of the marker, the date received and the date marked. Such a system should help the test provider to estimate the time and markers required for a given number of test takers.

The tracking system can also provide basic information to appraise the performance of each marker, such as the average time they take to mark a paper. If the marker's work is checked, the average number of errors they make can also be calculated. It may only be necessary to check a representative sample of each marker's work to produce these statistics.

5.1.2 Machine marking

Marking papers by machine usually implies the use of OPTICAL MARK READING/optical mark recognition (OMR) technology. OMR is most useful when large numbers of papers are to be marked, and when items are of types that do not require any human judgement (e.g. multiple-choice, true/false or multiple-matching types). Test takers can record their answers on customised OMR sheets, as illustrated in Figure 13, which can be fed into an OMR scanner so that the response data is captured and stored on computer. OMR technology can also be used for items which require clerical marking. The clerical marker records marks on the OMR sheet and it is then scanned.

Scanners speed up the process of marking, and reduce human error, but the scanning process may fail to read marks on the paper, or read unintended marks in error. In order to avoid such mistakes, integrity checks may be run on the data. This involves reviewing the OMR sheets to find apparent responses which are counter to test requirements, such as more than one selection for an item where only one is required. Any corrections will then need to be made to the OMR sheets by hand.

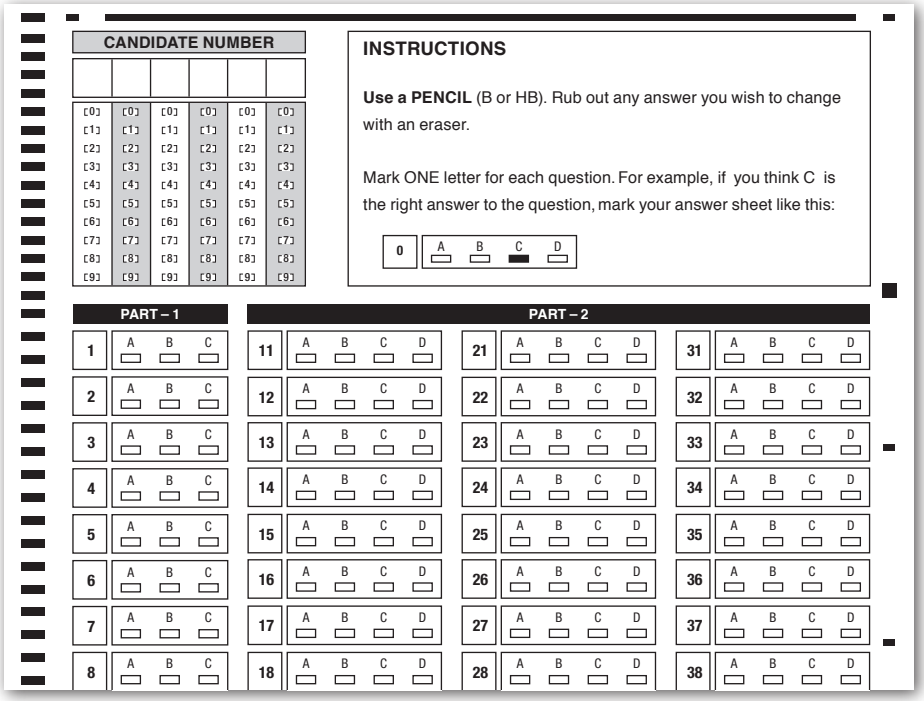


Figure 13 ***Section from an OMR sheet***

5.1.3 Rating

We will use *rating* and *raters* here to refer to marking where the exercise of trained judgement is necessary, to a much greater degree than in clerical marking. When judgement is used, a single 'correct answer' cannot be clearly prescribed by the exam provider before rating. For this reason, there is more scope for disagreement between judgements than in other kinds of marking, and thus a greater danger of inconsistency, between raters, or in the work of an individual rater. A combination of training, monitoring and corrective feedback may be used to ensure rating is accurate and reliable.

Much of what has been said of clerical marking is also true of rating: the process should be managed to use resources effectively, and checks and monitoring should help ensure accuracy. Reliability of ratings should also be monitored (see Section 1.3, Appendix VII).

Rating scales

Most approaches to rating proficiency depend on some kind of RATING SCALE. This is a set of descriptors which describe performances at different levels, showing which mark or grade each performance level should receive.

Rating scales reduce the variation inherent in the subjectivity of human judgements. There are a range of options to consider:

- **Holistic or analytic scales:** A single mark for a performance can be given using a single scale describing each level of performance, perhaps in terms of a range of features. The rater chooses the level which best describes the performance. Alternatively scales can be developed for a range of criteria (e.g. communicative effect, accuracy, coverage of expected content, etc.), and a mark given for each of these. Both approaches may relate to a similar language proficiency construct described in similar words – the difference is in the judgement the rater is required to make.
- **Relative or absolute scales:** Scales may be worded in relative, evaluative terms (e.g. 'poor', 'adequate', 'good'), or may aim to define performance levels in positive, definite terms. If we wish to interpret performance in terms of CEFR scales and levels, this second option seems preferable, and the CEFR descriptor scales are a useful source for constructing such rating scales.
- **Scales or checklists:** An alternative or additional approach to rating against a scale is to give marks based on a list of yes/no judgements as to whether a performance fulfils specific requirements or not.
- **Generic or task-specific scales:** An exam may use a generic scale or set of scales for all tasks, or provide rating criteria which are specific to each task. A combination of both is possible: for example, specific criteria may be provided for rating task fulfilment (a list of content points that should be addressed), while other scales are generic.
- **Comparative or absolute judgement:** It is possible to define a scale through exemplar performances, such that the rater's task is not to judge the absolute level of a performance, but simply to say whether it is lower, higher or the same in relation to one or more exemplars. A mark is thus a ranking on a scale. Interpretation of this ranking, e.g. in terms of CEFR levels, then depends on a judgement about the level represented by the exemplars. This approach will probably work best if the exemplars are task specific.

While these approaches may appear to differ widely, they all depend on similar underlying principles:

- all rating depends on the raters understanding levels
- exemplars are essential to defining and communicating this understanding
- the test tasks used to generate the rated performance are critically important to working with scales.

Traditionally levels had meaning which tended to be local to the context of a specific exam and its candidature, and thus difficult to compare with exam levels from other contexts. The development of proficiency frameworks such as the CEFR offers the possibility for levels used in local contexts to be understood in relation to other contexts. This impacts on the way rating scales are articulated.

If traditionally the level was implicit and understood, so that scales tended to be framed in relative, evaluative terms, nowadays scales are more likely to be framed in terms which echo the CEFR's approach of describing performance levels in recognisable ways, using positive and definite statements. This doesn't change the fact that exemplars (rather than the written text of the descriptors) remain essential to defining and

communicating the level, but it is good in that it encourages exam providers to be more explicit about what achieving a level means.

The CEFR promotes thinking and working in terms of criterion proficiency levels. There are two aspects to defining levels: *what* people can do and *h*ow well they can do them. In an exam the what is defined through the tasks which are specified. How well these tasks are done is what raters have to judge.

This is why the traditional evaluative approach to framing rating scales works well enough, as long as the tasks are appropriately chosen and the judgements relate to performance on the tasks. Thus tasks are critically important to defining scales, even if they may be more or less explicitly referred to in defining what a 'passing' performance means.

The CEFR (p. 188) discusses aspects of subjective assessment.

The rating process

For rating to work well raters must have a shared understanding of the standard. The basis of this shared understanding is shared examples of performance.

For small-scale exams a group of raters may arrive at a shared understanding through free and equal discussion. This may mean that test takers are treated equally, but it does not guarantee that the standard agreed will have more than local meaning or be stable across sessions. For large-scale exams the standard must be stable and meaningful. This will depend in practice on experienced examiners who have authority to communicate the standard to newcomers.

Thus a small group of experienced raters tends to act as the core maintaining the standard: training, monitoring and correcting other raters.

Such a hierarchical system can have several levels, as illustrated in Figure 14. This may be an effective way of managing face-to-face training, or the monitoring of markers' work. To some extent, however, modern information technology, and the development of web-based training, reduces the need for such a hierarchy. Also it is important to note that the same authoritatively marked exemplars at each level ensure accurate transmission of the standard.

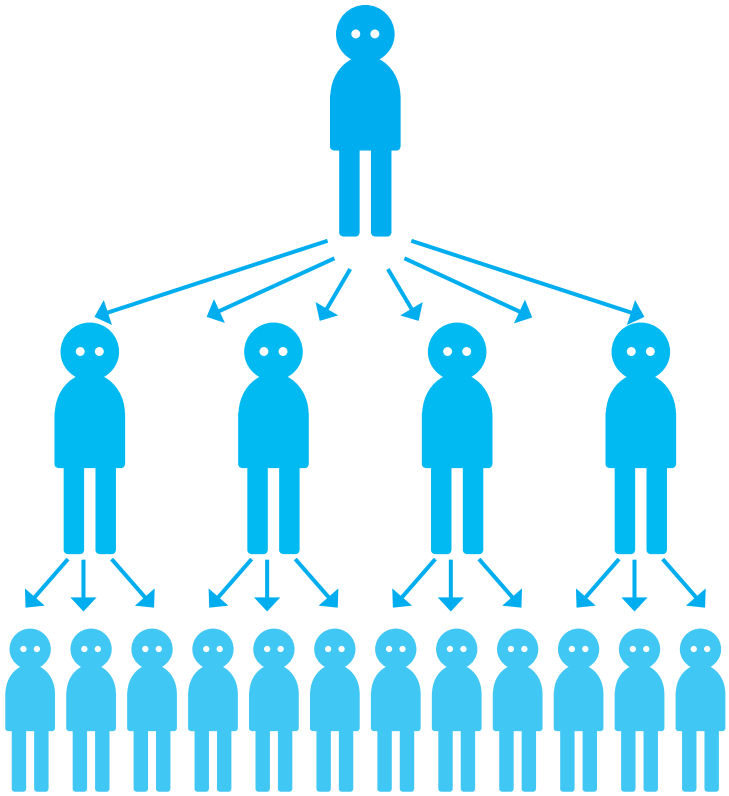


Figure 14 ***Maintaining standards through a team leader system***

Rater training

Rater training aims at consistent and accurate marking. *Standardisation* is the process of training raters to apply the intended standard. If the CEFR is the reference point for standards, then training should begin with exercises to familiarise raters with the CEFR, and may include reference to CEFR-related illustrative samples of performance for speaking or writing (Council of Europe 2009). It may also be necessary to 'train raters out of' using a rating scale which they are already familiar with. Training should then proceed through a series of steps from more open discussion towards independent rating, where the samples used relate to the exam being marked:

- guided discussion of a sample, through which markers come to understand the level
- independent marking of a sample followed by comparison with the pre-assigned mark and full discussion of reasons for discrepancies
- independent marking of several samples to show how close markers are to the pre-assigned marks.

Where possible, samples should represent performance on tasks from the current test session. If this is not possible, then tasks from previous sessions should be used.

Monitoring and quality control

Ideally the training phase ends with all markers working sufficiently accurately and consistently to make further feedback or correction unnecessary. This allows the marking phase to continue as smoothly as possible. However, some monitoring is necessary to identify problems early enough to remedy them.

We can identify four types of problem, or 'rater effects':

1. **Severity or lenience:** the rater systematically marks too low or too high.
2. **Use of mark range (central tendency):** the rater uses too narrow a range of marks, so does not distinguish high and low performance clearly enough.
3. **Halo effect:** where the rater has to assign several marks, they form an impression of the test taker from the first mark, and apply it to all the other marks, irrespective of the actual level of performance.
4. **Inconsistency:** the rater is unsystematic in the standard applied, so that their marks do not agree with other raters.

How serious these problems are depends partly on what kinds of correction are possible. Regarding severity, it appears that many raters have an inbuilt level of severity, and attempts to standardise this out of existence can backfire, by reducing the markers' confidence and making them less consistent. Thus it may be preferable to accept a level of systematic severity or leniency as long as there is a statistical procedure available to correct it. Scaling, or using an item response model, are two options here (see Appendix VII).

Use of too narrow a mark range can only be partially corrected statistically. Inconsistency is impossible to correct statistically, so both of these problems need to be identified and remedied, by either retraining the marker or dismissing them.

Some system of monitoring is thus desirable, although it is more easily done in real time in the case of writing, where a script may be passed from rater to rater. Oral assessment is much more difficult to monitor unless recorded. In this case more effort should be devoted to training and appraisal of the rater prior to the session. Statistics on the performance of the rater may be generated to help with this process (see Appendix VII).

Approaches to monitoring range from the simple – e.g. informal spot checking and provision of oral feedback to markers – to the complex – e.g. partial re-marking of a marker's work and generation of statistical indices of performance. An attractive method is to include pre-marked scripts in a marker's allocation and observe how closely their marks agree. However, for this to work reliably these scripts should be indistinguishable from other scripts, so photocopying, for example, is not possible. Practically this method is feasible only for scripts elicited in a computer-based test, or where paper scripts are scanned into an online marking system.

Another way of reducing rating error, and also of comparing raters with each other (which enables some rater effects to be identified and corrected statistically) is to use DOUBLE MARKING or *partial multiple marking*, where a proportion of scripts are marked by more than one rater. Depending on the statistical approach used, some method is needed to combine the information and arrive at a final mark for the test taker.

5.2 Grading

The whole process of test design, development, administration and marking which has been described so far leads up to the point where we can evaluate the performance of each test taker and report this in some way.

In some contexts a test simply ranks test takers from highest to lowest, perhaps setting arbitrary grade boundaries that divide them into groups – e.g. the top 10% get Grade A, the next 30% get Grade B and so on. This *norm-referenced* approach, though it may serve important purposes in society, is unsatisfactory to the extent that performance is evaluated only relative to other test takers – it says nothing about what performance *means* in terms, perhaps, of some useful level of language competence.

The alternative, more meaningful approach is *criterion-referenced* where performance is evaluated with respect to some fixed, absolute criterion or standard. Clearly, this is the case with language tests that report results in terms of CEFR levels.

An exam may be designed to report over several CEFR levels, or just one. In the latter case, those test takers who achieve the level may be said to have 'passed', and the others to have 'failed'. Degrees of passing or failing performance may also be indicated.

Identifying the score which corresponds to achieving a certain level is called STANDARD SETTING. It inevitably involves subjective judgement, as far as possible based on evidence.

Different approaches to standard setting apply to performance skills (speaking, writing) and to receptive skills (reading, listening) which are often objectively marked. The performance skills are much easier to address. Reading and listening are harder, because we have to interpret mental processes that are only indirectly observable, so that the notion of a criterion level of competence is hard to pin down.

Where a test comprises several different skill components a standard must be set for each skill separately, leaving the problem of how to summarise the results (see Section 5.3 for more on this).

The reader is referred to the *Manual for Relating Language Examinations to the CEFR* (Council of Europe 2009), which contains extensive treatment of standard setting. With respect to the organisation and terminology of the Manual, note that:

- Chapter 6 on *Standard Setting Procedures* refers to objectively marked tests only, i.e. reading, listening.
- Performance skills are treated under the heading of *Standardisation Training and Benchmarking* in Chapter 5.
- Chapter 7 on VALIDATION should also be carefully considered. There are two broad approaches to standard setting: task centred and learner centred. Task-centred approaches depend on expert judgements about test items, and this is what is treated in Chapter 6. Learner-centred approaches look to collecting additional evidence about learners from beyond the test. This is treated in Chapter 7.
- This organisation should not be taken to mean that task-based STANDARD SETTING is more important than learner-centred approaches.

Strictly speaking, standard setting is something that should happen once, when the exam is first administered, although in practice hitting the desired standard may be an iterative process. Over time, however, we would wish to see grading as a matter, not of setting, but of *maintaining* standards. This implies having appropriate procedures in place throughout the test development cycle. This is discussed in the materials additional to the Manual (North and Jones 2009).

5.3 Reporting of results

The user must decide whether a single result should be reported for each test taker, or a results profile for each test taker, showing performance on each test component.

The first of these is most common, reflecting the fact that most end users of exam results seem to prefer a simple answer to a complex one. The second provides a more informative report which might be very useful for some purposes.

A third option is to provide both. The CEFR stresses the importance of reporting profiled scores where possible.

Where a single result is required a method must be chosen for aggregating the marks for each skill component. The user must decide how each component should be WEIGHTED: all the skills equally, or some more than others. This may require some adjustment of the RAW SCORES on each component. See Appendix VII.

Where a results certificate is provided the user must consider:

- what additional material (e.g. 'Can Do' statements) might be provided to illustrate the meaning of the level
- whether the certificate should be guaranteed as genuine in some way (e.g. making forgery or alteration difficult, or providing a verification service)
- what caveats, if any, should be stated about interpretations of the result.

5.4 Key questions

- How much clerical marking will your test require and how often?
- How much rating will your test require and how often?
- What level of expertise will raters require?
- How will you ensure marking and rating is accurate and reliable?
- What is the best way to grade test takers in your context?
- Who will you report results to and how will you do it?

5.5 Further reading

See ALTE (2006c) for a self-assessment checklist for marking, grading and results.

Kaftandjieva (2004), North and Jones (2009) and Figueras and Noijons (2009) all provide information on standard setting.

6 Monitoring and review

It is important to check the work done to develop and use the test. Is it of an acceptable standard, or are changes necessary? The aim of monitoring is to verify that important aspects of a test are acceptable during or soon after the time it is used. If changes are necessary, it is often possible to make them quickly. Improvements can benefit the current test takers, or those taking the test next time.

Test review is a kind of project which looks at many aspects of the test. It also goes back to test development and asks fundamental questions, such as 'is this test needed?', 'for what purpose?', 'who is it for?' and 'what are we trying to test?'. It is like the test development stage but with the advantage of data and experience of using the test many times. Because of its size and scope, test review cannot be part of the normal testing cycle and cannot take place every session.

6.1 Routine monitoring

Monitoring is part of the routine operation of producing and using the test. Evidence gathered for monitoring will be used to ensure that everything concerned with the current test form is on track: materials are properly constructed, that they are delivered on time, test takers are given the correct grades, etc. After that, the same evidence can be used to appraise the performance of the processes used, such as the item writing and editing processes, TEST CONSTRUCTION process, the marking process, etc. The evidence may also be relevant to the VALIDITY ARGUMENT (see Appendix I), and should also be reviewed with this in mind.

Several examples of collecting evidence for monitoring are already described in this Manual. Examples of monitoring include:

- using expert judgement and TRIALLING or PRETESTING to ensure items are well written (see Section 3.4)
- using test taker responses to decide if items work appropriately (Appendix VII)
- using feedback forms to see if administration went well (see Appendix VI)
- collecting and analysing data about marker performance (see Appendix VII).

Monitoring the efficiency of the work may also be important. Test providers can measure how long a stage of preparation takes and decide that too much or too little time was set aside.

6.2 Periodic test review

Periodic reviews are done occasionally, outside of the regular operation of the test. This may be after a certain regular period, or after important changes to the circumstances of the test, such as if the target group of test takers or the use of the test changes, or the related syllabus changes. It is also possible that the need for review was spotted during monitoring. Reviews allow an extensive consideration of the test and the way in which it is produced. Evidence collected during the use of the test, such as while monitoring rater performance, may be useful in a review. In addition, the test provider may decide that other evidence is needed and this can be gathered especially for the review.

During a review information about the test is gathered and collated. This information can help decide what aspects of the test (e.g. the construct, the format, the regulations for administration) must be dealt with. It may be that the review recommends limited or no changes.

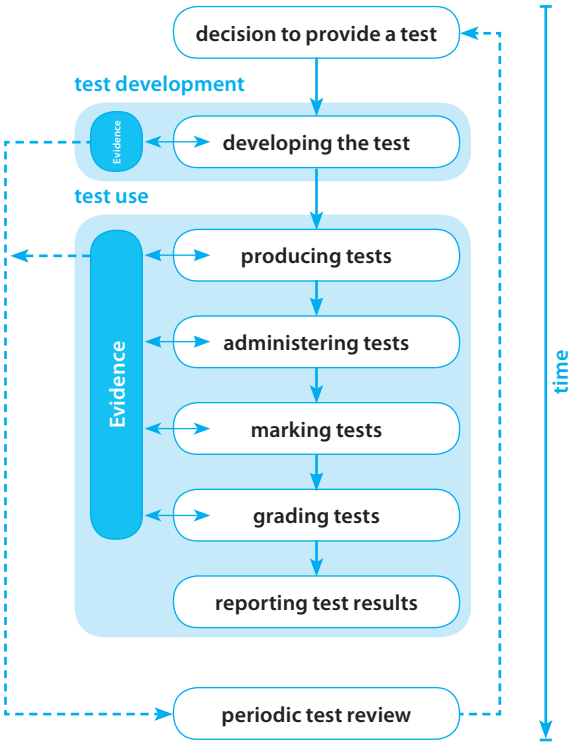


Figure 15 ***The testing cycle and periodic test review***

Figure 15 is a reproduction of Figure 5 (Section 1.5.1) with the addition of periodic review. It shows that the findings of the review feed into the very first stage in the diagram: the decision to provide a test. The test development process is completed again as part of the review.

Care should be taken to notify STAKEHOLDERS and others of changes, as suggested in Section 2.6.

6.3 What to look at in monitoring and review

Monitoring and review are parts of routine test development and use. They show the test provider whether everything is working as it should be and what to change if it is not. Review can also help to show others, such as school governors or accreditation bodies, that they can trust the exam. In both cases, finding out what is done and whether it is good enough is rather like an audit of the VALIDITY ARGUMENT.

ALTE (2007) has produced a list of 17 key points, called *Minimum Standards*, which allow test providers to structure their validity argument. They are divided into the following five general areas:

- test construction
- administration and logistics
- marking and grading
- test analysis
- communication with stakeholders.

They may be used with more detailed and specific lists, such as the ALTE Content Analysis Checklists (ALTE 2004a–k, 2005, 2006a–c).

Other tools are available to help test providers construct and check their validity arguments. Jones, Smith and Talley (2006:490–2) provide a list of 31 key points for small-scale achievement testing. Much of their list is based on *Standards for Educational and Psychological Testing* (AERA et al 1999).

6.4 Key questions

- What data needs to be collected to monitor the test effectively?
- Is some of this data already collected to make routine decisions during test use? How can it be used easily for both purposes?
- Can the data be kept and used later in test review?
- Who should be involved in test review?
- What resources are available for test review?
- How often should test review take place?
- Could any of the lists of key points be useful for checking the validity argument?

6.5 Further reading

ALTE (2007) provides a number of categories through which to assess a test.

See ALTE (2002) for a self-assessment checklist for test analysis and review.

Fulcher and Davidson (2009) illustrate an interesting way to think about using evidence for exam revision. They use the metaphor of a building to consider those parts of a test which must be changed more regularly and those which may be changed infrequently.

Descriptions of various aspects of test revision can be found in Weir and Milanovic (2003).

Bibliography and tools

AERA, APA, NCME (1999) *Standards for Educational and Psychological Testing*, Washington DC: AERA Publishing.

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

ALTE (1994) *Code of Practice*. Website. Access date: 12/07/09. Available at: **http://www.alte.org/downloads/index.php**

ALTE (2002) *ALTE Quality Management and Code of Practice Checklist*: *test analysis and post examination review*. Access date: 12/07/09. Available at: **http://www.alte.org/cop/copcheck.php**

ALTE (2004a) *Development and descriptive checklist for tasks and examinations*: *general*. Access date: 12/07/09. Downloaded from: **http://www.alte.org/downloads/index.php**

ALTE (2004b) *Individual component checklist*: *reading*. Access date: 12/07/09. Downloaded from: **http://www.alte.org/downloads/index.php**

ALTE (2004c) *Individual component checklist*: *structural competence*. Access date: 12/07/09. Downloaded from: **http://www.alte.org/downloads/index.php**

ALTE (2004d) *Individual component checklist: listening*. Access date: 12/07/09. Downloaded from: **http://www.alte.org/downloads/index.php**

ALTE (2004e) *Individual component checklist*: *writing*. Access date: 12/07/09. Downloaded from: **http://www.alte.org/downloads/index.php**

ALTE (2004f) *Individual component checklist*: *speaking*. Access date: 12/07/09. Downloaded from: **http://www.alte.org/downloads/index.php**

ALTE (2004g) *Individual component checklist – for use with one task*: *reading*. Access date: 12/07/09. Downloaded from: **http://www.alte.org/downloads/index.php**

ALTE (2004h) *Individual component checklist – for use with one task*: *structural competence*. Access date: 12/07/09. Downloaded from: **http://www.alte.org/downloads/index.php**

ALTE (2004i) *Individual component checklist – for use with one task*: *listening*. Access date: 12/07/09. Downloaded from: **http://www.alte.org/downloads/index.php**

ALTE (2004j) *Individual component checklist – for use with one task*: *writing*. Access date: 12/07/09. Downloaded from: **http://www.alte.org/downloads/index.php**

ALTE (2004k) *Individual component checklist – for use with one task: speaking*. Access date: 12/07/09. Downloaded from: **http://www.alte.org/downloads/index.php**

ALTE (2005) *ALTE materials for the guidance of test item writers* (*1995, updated July 2005*). Accessed: 12/07/09. Downloaded from: **http://www.alte.org/downloads/index.php**

ALTE (2006a) *ALTE Quality Management and Code of Practice Checklist*: *test construction*. Access date: 12/07/09. Available at: **http://www.alte.org/cop/copcheck.php**

ALTE (2006b) *ALTE Quality Management and Code of Practice Checklist*: *administration and logistics*. Access date: 12/07/09. Available at: **http://www.alte.org/cop/copcheck.php**

ALTE (2006c) *ALTE Quality Management and Code of Practice Checklist: marking, grading, results*. Access date: 12/07/09. Available at: **http://www.alte.org/cop/copcheck.php**

ALTE (2007) *Minimum standards for establishing quality profiles in ALTE Examinations*. Access date: 12/07/09. Downloaded from: **http://www.alte.org/downloads/index.php**

ALTE (2008a) *The ALTE Can Do Project*. Website. Access date: 12/07/09. Downloaded from: **http://www.alte.org/downloads/index.php**

ALTE (2008b) *ALTE Quality Management and Code of Practice Checklists*. Webpage. Access date: 12/07/09. Available at: **http://www.alte.org/cop/copcheck.php**

ALTE Members (1998) *Multilingual glossary of language testing terms* (Studies in Language Testing volume 6), Cambridge: Cambridge University Press.

ALTE Members (2005a) *The CEFR Grid for Speaking, developed by ALTE Members (input) v. 1.0.* Access date: 04/03/09. Downloaded from: **http://www.coe.int/T/DG4/Portfolio/documents/ALTE%20CEFR%20Speaking%20Grid%20INput51.pdf**

ALTE Members (2005b) *The CEFR Grid for Speaking, developed by ALTE Members (input) v. 1.0.* Access date: 04/03/09. Downloaded from: **http://www.coe.int/T/DG4/Portfolio/documents/ALTE%20CEFR%20Speaking%20Grid%20OUTput51.pdf**

ALTE Members (2007a) *The CEFR Grid for Writing Tasks v. 3.1 (analysis)*. Access date: 04/03/09. Downloaded from: **http://www.coe.int/T/DG4/Portfolio/documents/CEFRWritingGridv3_1_analysis.doc**

ALTE Members (2007b) *The CEFR Grid for Writing Tasks v. 3.1 (presentation)*. Access date: 04/03/09. Downloaded from: **http://www.coe.int/T/DG4/Portfolio/documents/CEFRWritingGridv3_1_presentation.doc**

ALTE Working Group on Code of Practice (2001) *The Principles of Good Practice for ALTE Examinations*. Access date: 12/07/09. Downloaded from: **http://www.alte.org/downloads/index.php**

Assessment Systems (2009) Iteman 4. Software. Assessment Systems.

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

Bachman, L F (2004) *Statistical Analysis for Language Assessment*, Cambridge: Cambridge University Press.

Bachman, L F (2005) Building and supporting a case for test use, *Language Assessment Quarterly* 2 (1), 1–34.

Bachman, L F; Black, P; Frederiksen, J; Gelman, A; Glas, C A W; Hunt, E; McNamara, T and Wagner, R K (2003) Commentaries Constructing an Assessment Use Argument and Supporting Claims About Test Taker-Assessment Task Interactions in Evidence-Centered Assessment Design, *Measurement: Interdisciplinary Research & Perspective* 1 (1), 63–91.

Bachman, L F and Palmer, A S (1996) *Language testing in practice: designing and developing useful language tests*. Oxford: Oxford University Press.

Bachman, L F and Palmer, A S (2010) *Language assessment in practice: developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Banerjee, J (2004) *Reference Supplement to the preliminary pilot version of the Manual for Relating Language examinations to the CEF: section D: Qualitative Analysis Methods*. Access date: 09/06/10. Available at: **http://www.coe.int/t/dg4/linguistic/CEF-ref-supp-SectionD.pdf**

Beacco J-C and Porquier, R (2007) *Niveau A1 pour le français. Un référentiel*, Paris: Editions Didier.

Beacco J-C and Porquier, R (2008) *Niveau A2 pour le français. Un référentiel*, Paris: Editions Didier.

Beacco, J-C; Bouquet, S and Porquier, R (2004) *Niveau B2 pour le français. Un référentiel*, Paris: Editions Didier.

Bolton, S; Glaboniat, M; Lorenz, H; Perlmann-Balme, M and Steiner, S (2008) *Mündliche – Mündliche Produktion und Interaktion Deutsch: Illustration der Niveaustufen des Gemeinsamen europäischen Referenzrahmens*, München: Langenscheidt.

Bond, T G and Fox, C M (2007) *Applying the Rasch model: fundamental measurement in the human sciences*, Mahwah, NJ: Lawrence Erlbaum.

Brennan, R L (1992) Generalizability Theory, *Instructional Topics in Educational Measurement Series 14*. Access date: 05/03/09. Downloaded from: **http://www.ncme.org/pubs/items/21.pdf**

Briggs, D C; Haertel, E; Schilling, S G; Marcoulides, G A and Mislevy, R J (2004) Comment: Making an Argument for Design Validity Before Interpretive Validity, *Measurement: Interdisciplinary Research & Perspective* 2 (3), 171–191.

Camilli, G and Shepard, L A (1994) *Methods for Identifying Biased Test Items*, Thousand Oaks, CA: Sage.

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

Carr, N T (2008) Using Microsoft Excel® to Calculate Descriptive Statistics and Create Graphs, *Language Assessment Quarterly* 5 (1), 43.

CEFTrain (2005) *CEFTrain*. Website. Access date: 04/03/09.
Available at: **http://helsinki.fi/project/ceftrain/index.html**

Chapelle, C A; Enright, M K and Jamieson, J M (2007) *Building a Validity argument for the Test of English as a Foreign Language*, Oxford: Routledge.

CIEP (2009) *Productions orales illustrant les 6 niveaux du Cadre européen commun de référence pour les langues*. Website. Access date: 12/07/09. Available at: **www.ciep.fr/publi_evalcert**

CIEP/Eurocentres (2005) *Exemples de productions orales illustrant, pour le français, les niveaux du Cadre européen commun de référence pour les langues*, DVD, Strasbourg: Council of Europe.

Cizek, G J (1996) Standard-setting guidelines, *Instructional Topics in Educational Measurement Series*.
Access date: 05/03/09. Downloaded from: **http://www.ncme.org/pubs/items/Standard.pdf**

Cizek, G J and Bunch, M B (2006) *Standard Setting: A Guide To Establishing And Evaluating Performance Standards On Tests*, Thousand Oaks, CA: Sage.

Cizek, G J; Bunch, M B and Koons, H (2004) *Setting performance standards: contemporary methods, Instructional Topics in Educational Measurement Series*. Access date: 05/03/09. Downloaded from:
http://www.ncme.org/pubs/items/Setting%20Performance%20Standards%20ITEMS%20Module.pdf

Clauser, B E and Mazor, K M (1998) Using statistical procedures to identify differentially functioning test items, *Instructional Topics in Educational Measurement Series*. Access date: 05/03/09. Downloaded from:
http://www.ncme.org/pubs/items/Statistical.pdf

Cook, L L and Eignor, D R (1991) IRT Equating Methods, *Instructional Topics in Educational Measurement Series 10*.
Access date: 05/03/09. Downloaded from: **http://www.ncme.org/pubs/items/17.pdf**

Corrigan, M (2007) *Seminar to calibrate examples of spoken performance, Università per Stranieri di Perugia, CVCL (Centro per la Valutazione e la Certificazione Linguistica) Perugia, 17th–18th December 2005*. Access date: 07/03/10.
Downloaded from: **http://www.coe.int/T/DG4/Portfolio/documents/Report_Seminar_Perugia05.pdf**

Coste, D (2007) *Contextualising Uses of the Common European Framework of Reference for Languages*, paper presented at Council of Europe Policy Forum on use of the CEFR, Strasbourg 2007; Downloaded from:
http://www.coe.int/t/dg4/linguistic/Source/SourceForum07/D-Coste_Contextualise_EN.doc

Council of Europe (1996) *Users' Guide for Examiners*, Strasbourg: Language Policy Division.

Council of Europe (1998) *Modern Languages: learning, teaching, assessment. A Common European Framework of Reference*, Strasbourg: Language Policy Division.

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, teaching, assessment*, Cambridge: Cambridge University Press.

Council of Europe (2004a) *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Access date: 04/03/09. Downloaded from:
http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf

Council of Europe (2005) *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR).* Reading and Listening Items and Tasks: Pilot Samples illustrating the common reference levels in English, French, German, Italian and Spanish, CD, Strasbourg: Council of Europe.

Council of Europe (2006a) *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*. Writing samples. Access date: 04/03/09. Downloaded from:
http://www.coe.int/T/DG4/Portfolio/documents/exampleswriting.pdf

Council of Europe (2006b) *TestDaF Sample Test Tasks*. Access date: 04/03/09. Downloaded from: **http://www.coe.int/T/DG4/Portfolio/documents/ALTECEFR%20Writing%20Grid-2.0_TestDaF%20samples.pdf**

Council of Europe (2009) *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR) – A Manual*. Access date: 04/03/09. Downloaded from: **http://www.coe.int/t/dg4/linguistic/Source/Manual%20Revision%20-%20proofread%20-%20FINAL.pdf**

Council of Europe and CIEP (2009) *Productions orales illustrant les 6 niveaux du Cadre européen commun de référence pour les langues*, DVD, Strasbourg and Sèvres: Council of Europe and CIEP.

Davidson, F and Lynch, B K (2002) *Testcraft: A teacher's guide to writing and using language test specifications*. New Haven: Yale University Press.

Davies, A (1997) (Guest Ed.) Ethics in language testing, *Language Testing* 14 (3).

Davies, A (2004) (Guest Ed.) *Language Assessment Quarterly* 2 & 3.

Davies, A (2010) Test fairness: a response, *Language Testing* 27 (2), 171-176.

Davies, A; Brown, A; Elder, C; Hill, K; Lumley, T and McNamara, T (1999) *Dictionary of language testing* (Studies in Language Testing volume 7), Cambridge: Cambridge University Press.

Downing, S M and Haladyna, T M (2006) *Handbook of Test Development*, Mahwah, NJ: Lawrence Erlbaum.

EALTA (2006) *EALTA Guidelines for Good Practice in Language Testing and Assessment*. Access date: 05/03/09. Downloaded from: **http://www.ealta.eu.org/documents/archive/guidelines/English.pdf**

Eckes, T (2009) *Reference Supplement to the preliminary pilot version of the Manual for Relating Language examinations to the CEF: section H*: Many-Facet Rasch Measurement. Access date: 09/06/10. Available at: **http://www.coe.int/t/dg4/linguistic/Source/CEF-refSupp-SectionH.pdf**

Education Testing Services (2002) *ETS Standards for Quality and Fairness*, Princeton, NJ: ETS.

Embretson, S E (2007) Construct Validity: A Universal Validity System or Just Another Test Evaluation Procedure?, *Educational Researcher* 36 (8), 449.

Eurocentres and Federation of Migros Cooperatives (2004) *Samples of oral production illustrating, for English, the levels of the Common European Framework of Reference for Languages*, DVD, Strasbourg: Council of Europe.

Europarat; Council for Cultural Co-operation, Education Committee, Modern Languages Division; Goethe-Institut Inter Nationes u.a. (Hg.) (2001) *Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen*, Berlin, München: Langenscheidt.

Figueras, N; Kuijper, H; Tardieu, C; Nold, G and Takala, S (2005) *The Dutch Grid Reading/Listening*. Website. Access date: 04/03/09. Available at: **http://www.lancs.ac.uk/fss/projects/grid/**

Figueras, N and Noijons, J (eds) (2009) *Linking to the CEFR levels: Research perspectives*. CITO/Council of Europe. Access date: 10/01/10. Downloaded from: **http://www.coe.int/t/dg4/linguistic/EALTA_PublicatieColloquium2009.pdf**

Frisbie, D A (1988) Reliability of Scores From Teacher-Made Tests, *Instructional Topics in Educational Measurement Series 3*. Access date: 05/03/09. Downloaded from: **http://www.ncme.org/pubs/items/ITEMS_Mod_3.pdf**

Fulcher, G and Davidson, F (2007) *Language Testing and Assessment – an advanced resource book*, Abingdon: Routledge.

Fulcher, G and Davidson, F (2009) Test architecture, test retrofit, *Language Testing* 26 (1), 123–144.

Glaboniat, M; Müller, M; Rusch, P; Schmitz, H and Wertenschlag, L (2005) *Profile Deutsch – Gemeinsamer europäischer Referenzrahmen. Lernzielbestimmungen, Kannbeschreibungen, Kommunikative Mittel, Niveau A1–A2, B1– B2, C1–C2*, Berlin, München: Langenscheidt.

Goethe-Institut Inter Nationes; der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (KMK); der Schweizerischen Konferenz der Kantonalen Erziehungsdirektoren (EDK) und dem österreichischen Bundesministerium für Bildung, Wissenschaft und Kultur (BMBWK), (Hg.) (2001) *Gemeinsamer Europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen*, im Auftrag des Europarats, Rat für kulturelle Zusammenarbeit, deutsche Ausgabe, München: Langenscheidt.

Gorin, J S (2007) Reconsidering Issues in Validity Theory, *Educational Researcher* 36 (8), 456.

Grego Bolli, G (Ed.) (2008) *Esempi di Produzioni Orali – A illustrazione per l'italiano dei livelli del Quadro comune europeo di riferimento per le lingue*, DVD, Perugia: Guerra Edizioni.

Haertel, E H (1999) Validity arguments for High-Stakes Testing: in search of the evidence, *Educational Measurement*: *Issues and Practice* 18 (4), 5.

Hambleton, R K and Jones, R W (1993) Comparison of classical test theory and item response theory and their applications to test development, *Instructional Topics in Educational Measurement Series 16*. Access date: 05/03/09. Downloaded from: **http://www.ncme.org/pubs/items/24.pdf**

Harvill L M (1991) Standard error of measurement, *Instructional Topics in Educational Measurement Series 9*. Access date: 05/03/09. Downloaded from: **http://www.ncme.org/pubs/items/16.pdf**

Heaton, J B (1990) *Classroom Testing*, Harlow: Longman.

Holland, P W and Dorans, N J (2006) Linking and Equating, in Brennan, R L (Ed.) *Educational measurement* (*4th edition*), Washington, DC: American Council on Education/Praeger.

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

ILTA (2000) *ILTA Code of Ethics*. Web page. Access date: 05/03/09. Available at: **http://www.iltaonline.com/index.php?option=com_content&task=view&id=57&Itemid=47**

ILTA (2007) *ILTA Guidelines for Practice*. Web page. Access date: 05/03/09. Available at: **http://iltaonline.com/index.php?option=com_content&task=view&id=122&Itemid=133**

Instituto Cervantes (2007) *Plan curricular del Instituto Cervantes – Niveles de referencia para el español*, Madrid: Edelsa.

JCTP (1988) *Code of Fair Testing Practices in Education*. Access date: 12/08/09. Downloaded from: **http://www.apa.org/science/fairtestcode.html**

JLTA (not specified) *Code of Good Testing Practice*. Access date: 12/08/09. Downloaded from: **http://www.avis.ne.jp/~youichi/COP.html**

Jones, N and Saville, N (2009) European Language Policy: Assessment, Learning and the CEFR, *Annual Review of Applied Linguistics* 29, 51–63.

Jones, P; Smith, R W and Talley, D (2006) Developing Test Forms for Small-Scale Achievement Testing Systems, in Downing, S M and Haladyna, T M (Eds) *Handbook of Test Development*, Mahwah, NJ: Lawrence Erlbaum.

Jones, R L and Tschirner, E (2006) *A Frequency Dictionary of German – Core Vocabulary for Learners*, New York: Routledge.

Kaftandjieva, F (2004) *Reference Supplement to the preliminary pilot version of the Manual for Relating Language examinations to the CEF*: *section B*: *Standard Setting*. Access date: 09/06/10. Available at: **http://www.coe.int/t/dg4/linguistic/CEF-refSupp-SectionB.pdf**

Kane, M (2002) Validating High Stakes Testing Programs, *Educational Measurement*: *Issues and Practices* 21 (1), 31–41.

Kane, M (2004) Certification Testing as an Illustration of Argument-Based Validation, *Measurement*: *Interdisciplinary Research & Perspective* 2 (3), 135–170.

Kane, M (2006) Validation, in Brennan, R L (Ed.) *Educational measurement* (*4th edition*), Washington, DC: American Council on Education/Praeger.

Kane, M (2010) Validity and fairness, *Language Testing* 27(2), 177-182.

Kane, M; Crooks, T and Cohen, A (1999) Validating measures of performance, *Educational Measurement*: *Issues and Practice* 18 (2), 5–17.

Kolen, M J (1988) Traditional Equating Methodology, *Instructional Topics in Educational Measurement Series 6*. Access date: 05/03/09. Downloaded from: **http://www.ncme.org/pubs/items/11.pdf**

Kolen, M J (2006) Scaling and Norming, in Brennan, R L (Ed.) *Educational measurement* (*4th edition*), Washington, DC: American Council on Education/Praeger.

Kuijper, H (2003) *QMS as a Continuous Process of Self-Evaluation and Quality Improvement for Testing Bodies*. Access date: 12/07/09. Downloaded from: **http://www.alte.org/qa/index.php**

Kunnan, A J (2000a) Fairness and justice for all, in Kunnan, A J (Ed.) *Fairness and validation in language assessmen*t , Cambridge: Cambridge University Press, 1–13.

Kunnan, A J (2000b) *Fairness and Validation in Language Assessment*: *Selected papers from the 19th Language Testing Research Colloquium*, *Orlando*, *Florida* (Studies in Language Testing volume 9), Cambridge: Cambridge University Press.

Kunnan, A J (2004) Test Fairness, in Milanovic, M and Weir, C (Eds) *European Language Testing in a Global Context – Proceedings of the ALTE Barcelona Conference*, *July 2001* (Studies in Language Testing volume 18), Cambridge: Cambridge University Press.

Linacre, J M (2009) Facets 3.64.0. Software. Winsteps.com software.

Lissitz, R W and Samuelsen, K (2007a) A Suggested Change in Terminology and Emphasis Regarding Validity and Education, *Educational Researcher* 36 (8), 437.

Lissitz, R W and Samuelsen, K (2007b) Further Clarification Regarding Validity and Education, *Educational Researcher* 36 (8), 482.

Livingston, S (2004) *Equating Test Scores* (*Without IRT*). Access date: 12/07/09. Downloaded from: **http://www.ets.org/Media/Research/pdf/LIVINGSTON.pdf**

McNamara, T and Roever, C (2006) Fairness Reviews and Codes of Ethics, *Language Learning* 56 (S2), 129–148.

Messick, S (1989) Meaning and values in test validation: the science and ethics of assessment, *Educational Researcher: Issues and Practice* 18, 5–11.

Messick, S (1989) Validity, in Linn, R (Ed.) *Educational measurement*, 3rd edition, New York: Macmillan, 13–103.

Mislevy, R J (2007) Validity by Design, *Educational Researcher* 36 (8), 463.

Mislevy, R J; Steinberg, L S and Almond, R G (2003) Focus Article: On the Structure of Educational Assessments, *Measurement*: *Interdisciplinary Research & Perspective* 1 (1), 3–62.

Moss, P A (2007) Reconstructing Validity, *Educational Researcher* 36 (8), 470.

North, B and Jones, N (2009) *Relating Language Examinations to the Common European Framework of Reference for Languages*: *Learning*, *teaching*, *assessment* (*CEFR*) – *Further Material on Maintaining Standards across Languages*, *Contexts and Administrations by exploiting Teacher Judgment and IRT Scaling*. Access date: 04/03/09. Downloaded from: **http://www.coe.int/t/dg4/linguistic/Manual%20-%20Extra%20Material%20-%20proofread%20-%20FINAL.pdf**

Parkes, J (2007) Reliability as Argument, *Educational Measurement*: *Issues and Practice* 26(4): 2-10.

Perlmann-Balme, M and Kiefer, P (2004) *Start Deutsch. Deutschprüfungen für Erwachsene. Prüfungsziele*, *Testbeschreibung*, München, Frankfurt: Goethe-Institut und WBT.

Perlmann-Balme, M; Plassmann, S and Zeidler, B (2009) *Deutsch-Test für Zuwanderer. Prüfungsziele*, *Testbeschreibung*, Berlin: Cornelsen.

Saville, N (2005) Setting and monitoring professional standards: A QMS approach, *Research Notes* 22. Access date: 05/03/09. Downloaded from: **http://www.cambridgeesol.org/rs_notes/rs_nts22.pdf**

Sireci, S G (2007) On Validity Theory and Test Validation, *Educational Researcher* 36 (8), 477.

Spinelli, B and Parizzi, F (2010) *Profilo della lingua italiana. Livelli di riferimento del QCER A1, A2, B1 e B2,* Milan: RCS libri – Divisione education.

Spolsky, B (1981) Some ethical questions about language testing, in Klein-Braley, C and Stevenson, D (Eds) *Practice and problems in language testing*, Frankfurt: Verlag Peter Lang, 5–21.

Stiggins, R J (1987) Design and Development of Performance Assessment, *Instructional Topics in Educational Measurement Series 1*. Access date: 05/03/09. Downloaded from: **http://www.ncme.org/pubs/items/ITEMS_Mod_1_Intro.pdf**

Traub, R E and Rowley, G L (1991) Understanding reliability, *Instructional Topics in Educational Measurement Series 8*. Access date: 05/03/09. Downloaded from: **http://www.ncme.org/pubs/items/15.pdf**

Trim, J L M (2010) Plenary presentation at ACTFL-CEFR Alignment Conference, Leipzig, June 2010.

University of Cambridge ESOL Examinations (2004) *Samples of oral production illustrating, for English, the levels of the Common European Framework of Reference for Language*, DVD, Strasbourg: Council of Europe.

University of Cambridge ESOL Examinations/Council of Europe (2009a) *Common European Framework of Reference for Languages Examples of Speaking Test Performance at Levels A2 to C2*, DVD, Cambridge: University of Cambridge ESOL Examinations.

University of Cambridge ESOL Examinations/Council of Europe (2009b) *Common European Framework of Reference for Languages Examples of Speaking Test Performance at Levels A2 to C2*. Website. Access date: 02/09/09. Available at: **http://www.cambridgeesol.org/what-we-do/research/speaking-performances.html**

van Avermaet, P (2003) *QMS and The Setting of Minimum Standards: Issues of Contextualisation Variation between The Testing Bodies*. Access date: 12/07/09. Downloaded from: **http://www.alte.org/qa/index.php**

van Avermaet, P; Kuijper, H and Saville, N (2004) A Code of Practice and Quality Management System for International Language Examinations, *Language Assessment Quarterly* 1 (2 & 3), 137–150.

van Ek, J A and Trim, J (1990) *Waystage 1990*, Cambridge: Cambridge University Press.

van Ek, J A and Trim, J (1991) *Threshold 1990*, Cambridge: Cambridge University Press.

van Ek, J A and Trim, J (2001) *Vantage*, Cambridge: Cambridge University Press.

Verhelst, N (2004a) *Reference Supplement to the preliminary pilot version of the Manual for Relating Language examinations to the CEF: section C: Classical Test Theory*. Access date: 09/06/10. Available at: **http://www.coe.int/t/dg4/linguistic/CEF-refSupp-SectionC.pdf**

Verhelst, N (2004b) *Reference Supplement to the preliminary pilot version of the Manual for Relating Language examinations to the CEF: section E: Generalizability Theory*. Access date: 09/06/10. Available at: **http://www.coe.int/t/dg4/linguistic/CEF-ref-supp-SectionE.pdf**

Verhelst, N (2004c) *Reference Supplement to the preliminary pilot version of the Manual for Relating Language examinations to the CEF: section F: Factor Analysis*. Access date: 09/06/10. Available at: **http://www.coe.int/t/dg4/linguistic/CEF-ref-supp-SectionF.pdf**

Verhelst, N (2004d) *Reference Supplement to the preliminary pilot version of the Manual for Relating Language examinations to the CEF: section G: Item Response Theory*. Access date: 09/06/10. Available at: **http://www.coe.int/t/dg4/linguistic/CEF-ref-supp-SectionG.pdf**

Ward, A W and Murray-Ward, M (1994) Guidelines for Development of item banks, *Instructional Topics in Educational Measurement Series 17*. Access date: 05/03/09. Downloaded from: **http://www.ncme.org/pubs/items/25.pdf**

Weir, C J (2005) *Language testing and validation: An evidence-based approach*, Basingstoke: Palgrave Macmillan.

Weir, C J and Milanovic, M (2003) (Eds) *Continuity and Innovation: Revising the Cambridge Proficiency in English Examination 1913–2002* (Studies in Language Testing volume 15), Cambridge: Cambridge University Press.

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

Xi, X (2010) How do we go about investigating test fairness?, *Language Testing* 27(2): 147-170.

Websites

Association of Language Testers in Europe: **www.alte.org**

English Profile: **www.englishprofile.org**

European Association for Language Testing and Assessment: **http://www.ealta.eu.org/**

International Language Testing Association: **http://www.iltaonline.com/**

Language Policy Division, Council of Europe: **http://www.coe.int/t/dg4/linguistic/default_EN.asp**

Appendix I – Building a validity argument

This appendix introduces an approach to VALIDATION which involves building a VALIDITY ARGUMENT. It is more detailed than the outline provided in 1.2.3 and shows how the steps in the argument are not actually discrete and sequential but overlapping and interrelated.

Kane (2006), Kane, Crooks and Cohen (1999), Bachman (2005) and Bachman and Palmer (2010) describe validity arguments more fully. For this reason, validation is an ongoing process, adding more evidence and refinement to the validity argument as time goes by.

The focus of a validity argument is the interpretation and use of test results, following the definition of validity as 'the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests' (AERA et al 1999).

A validity argument is a series of propositions which describe why recommended interpretations of test results are valid and provide evidence and theory to support this. This appendix provides an overview of how to do this.

When the argument is presented to STAKEHOLDERS, the starting point is a clear statement outlining how test results should be interpreted for a particular use. An ASSESSMENT USE ARGUMENT (also called an INTERPRETIVE ARGUMENT) explains this statement. What is usually just called a validity argument sets out to justify the use argument with theory and evidence.

Figure 16 shows a conceptual view of a use argument after Bachman (2005). It is a chain of reasoning, in four steps (each shown by an arrow), which justifies the use of the test results. Each step provides a conceptual basis for the one following it. For example, reliable test scores (universe score) are only useful if they adequately represent test performance (observed score). The diagram does *not* represent a sequence of stages which must be completed one after the other, and evidence to support each step may be gathered from various stages in test development and production.

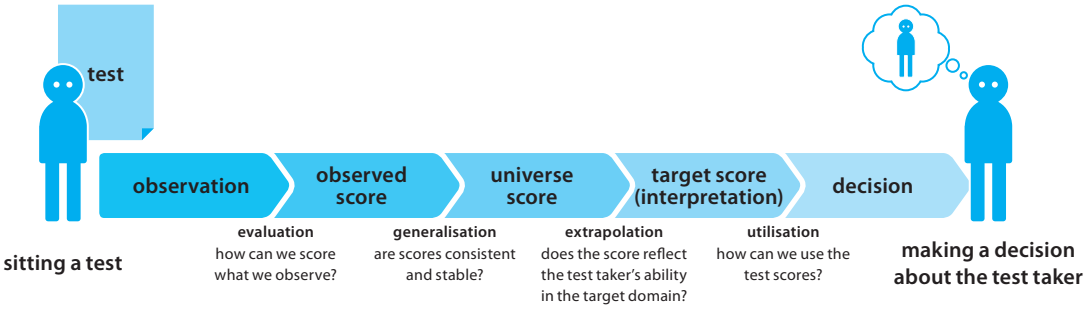


Figure 16 ***Chain of reasoning in a validity argument (adapted from Kane, Crooks, Cohen 1999, Bachman 2005)***

The supporting validity argument is constructed to support the use argument and consists of evidence, theory and reasoned propositions. The evidence which supports each step is gathered during test development, construction and use.

Much of the evidence for a validity argument will come from the routine process of using the test. Examples of this kind of evidence have been listed in Section 6.1. The evidence is collected for another, more immediate purpose, such as monitoring rater performance, but will also be useful when building a validity argument. This is illustrated in Figure 17.

The validity argument can be improved and developed using evidence each time a new test form is developed and used. The development of the validity argument should begin at a very early stage of the process, when the intended purposes of the test are defined. However, much of the validity argument for one test form can be used as the argument for the next.

Some theorists (Bachman 2005, Mislevy et al 2003) stress that a validity argument should take the form of an *informal argument* in contrast to a logical argument. This means that the argument cannot be proven right or wrong by reasoning alone. Instead it can seem more or less convincing to someone reviewing it. How convincing it is will depend on the theory and evidence available to support it.

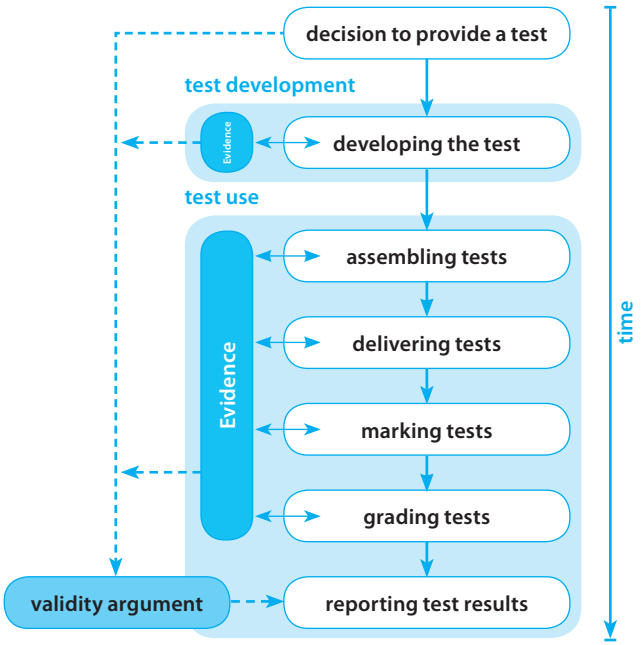


Figure 17 ***The testing cycle, periodic review and validity argument***

The validity argument could be made to seem less convincing by new evidence or theory, or a different interpretation of some existing evidence. There is also the chance that test providers will, unintentionally, simply confirm the interpretations they prefer without being critical enough. After the argument has been developed for the first time, therefore, it must be challenged, even if this could mean altering the recommended interpretation of test results. This may be done by, for example, considering alternative ways to understand the evidence or to check that all the inferences in the argument are sound. The test provider should then reconsider their argument, make any necessary changes and present reasons why all the evidence was interpreted as it was.

Examples of different evidence that can be used to support a validity argument are given in the remainder of this appendix. In addition, examples of different ways to understand the evidence are given. All these examples are based on the work of Kane (2004) and Bachman (2005). They are arranged to follow the structure of this Manual: Developing tests, Assembling tests, Delivering tests, and Marking, grading and the reporting of results. Test providers can consider this evidence as a starting point for the construction of their own validity arguments; the lists, however, are not exhaustive.

Further reading

ALTE (2005:19) provides a useful summary of types of validity and the background to the modern conception of validity.

AERA et al (1999) provide an outline of the modern concept of validity and standards which highlight specific areas of concern, and can therefore help in the development of a validity argument.

Messick (1989) discusses the unitary concept of validity and also ethical considerations which stem from it.

Haertel (1999) provides an example of how evidence and argumentation may be related to score interpretations.

Kane, Crooks and Cohen (1999) present a clear picture of the first stages in a validity argument. A more in-depth treatment is available in Kane (2006).

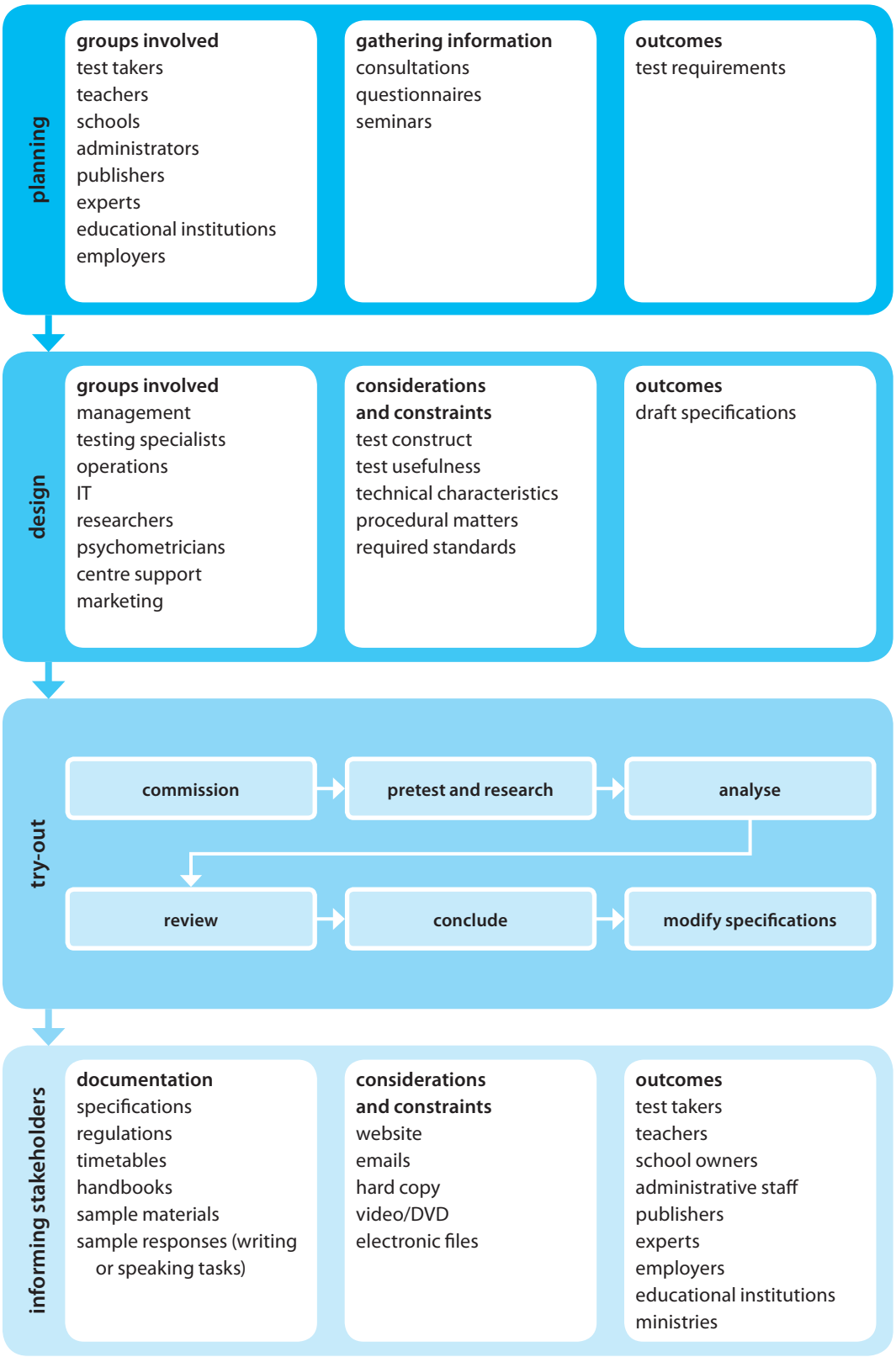
Bachman (2005) discusses validity arguments in relation to language assessment. He also maps Bachman and Palmer's (1996) model to the validity argument model. In the earlier model, usefulness was seen as the most significant quality of a test, being a balance between reliability, validity, authenticity, interactivity and impact.

Bachman and Palmer (2010) show how validity arguments are central to test development and can provide a framework for these tasks.

	evaluation	generalisation	extrapolation	utilisation
	how can we score what we observe?	**are scores consistent and stable?**	**does the score reflect the test taker's ability in the target domain?**	**how can we use the scores?**
TEST DEVELOPMENT (Section 2)				
evidence for		The specifications require a standard test format – this will support the idea that different test forms are similar to each other (see Section 2 and Appendix III).	The item writer guidelines and specifications clearly target a domain of use. This domain may also be identified in a needs analysis (see Section 2.4). Evidence that pass marks are appropriately set will support the recommended interpretation of each test taker's results (see Sections 2.0 and 5.2).	
evidence against			Some areas of the CONSTRUCT are not fully represented in the test specifications. This would mean that the test results would not provide adequate information about what the test takers can do (see Sections 1.1 and 2).	
TEST PRODUCTION (Section 3)				
evidence for	All marking keys are correct. Grammar books, dictionaries and expert knowledge can be used to verify this.	Items in one test form represent the construct just as much as the items in another form. It is unlikely that exactly the same areas are covered each time, but areas of the construct must be selected in a way which is comparable (see Sections 2, 3.5 and Appendix VII). The way of linking forms is appropriate (see Appendix VII). If statistical analysis was used, low levels of error were found and statistical MODELS FITTED the data (see Appendix VII).	Experts were used in item writing and test construction (see Section 3.2).	
evidence against		Test forms have not been linked. Test forms do not represent the same construct.	It is possible that some areas of the construct were not adequately represented in the test materials. This would mean that the results would not provide adequate information about what the test takers can do.	

	evaluation	generalisation	extrapolation	utilisation
TEST DELIVERY (Section 4)				
evidence for	Procedures were followed during test administration. This will help to show that the test score is not influenced by other factors (such as too much, or too little time) (see Section 4.2).	Procedures have always been followed and will help to show that test forms are comparable in different administrations (see Section 4.2).		
evidence against	Unaddressed cheating will mean that test scores do not represent the ability of the test taker.	Undetected cheating will mean the scores of some test takers will not sufficiently represent their language ability. This is likely to be different across test forms.	Outside factors could have affected the test results. This may be because administration procedures were not followed, for example. As a consequence, test results also reflect these outside factors and it is difficult to say that they only relate to language ability (see Section 4.1).	
MARKING, GRADING AND THE REPORTING OF RESULTS (Section 5)				
evidence for	Procedures were followed during marking. This will help to show that the test score is not influenced by other factors (such as use of the wrong key, or scanning errors) (see Section 5.0). Marking was accurate and reliable (see Section 5.1 and Appendix VII).	Evidence of score reliability (usually statistical evidence), can show that this test form measures test takers in a consistent way (see Sections 1.3, 5.1 and Appendix VII). If data from only some test takers was analysed, this data is representative of the entire group of test takers (See Appendix VII). Cut-off points that have low levels of error will mean that it is more likely that test takers are placed on the correct side of the grade boundaries (see Appendix VII).	The use of expert raters means that ratings are more likely to reflect the domain of interest (see Section 5.1). Similarly, the use of well-written rating scales will increase the chances that performances are rated according to the target domain (see Sections 2.5 and 5.1.3).	If rules are used to make specific decisions based on test results, it is more likely that the test will be used as planned and undesired effects will be minimised (see Sections 1.2, 5.3 and Appendix I).
evidence against		If data from an unrepresentative group of test takers was used for analysis, the analysis may contain error and/or bias (see Appendix VII).		If rules and standard procedures were not followed in making decisions, the test may be used inappropriately (see Sections 1.5 and 5.3).

Appendix II – The test development process



< Back to contents

Appendix III – Example exam format – English sample

Content and overview

Paper/Timing	Format		No. of Qs	Test focus
READING 1 hour	**Part 1**	A MATCHING TASK involving one continuous text divided into four sections or four informational texts; approximately 250–350 words in total.	7	An emphasis on scanning and reading for gist.
	Part 2	A matching task involving a single text (article, report, etc.) with sentence-length gaps; approximately 450–550 words.	5	Understanding text structure.
	Part 3	A 4-option multiple-choice task involving a single text; approximately 450–550 words.	6	Reading for gist and specific information.
	Part 4	A 4-option multiple-choice cloze involving a single informational text with lexical gaps; text including gapped words; approximately 200–300 words.	15	Vocabulary and structure.
	Part 5	A proofreading task involving identification of additional unnecessary words in a short text; approximately 150–200 words.	12	Understanding sentence structure and error identification.
WRITING 45 minutes	**Part 1**	A message, memo or email. Test takers are required to produce an internal communication based on a RUBRIC only (plus layout of output text type); 40–50 words.	One compulsory task	Giving instructions, explaining a development, asking for comments, requesting information, agreeing to requests.
	Part 2	Business correspondence, short report or proposal. Test takers are required to produce a piece of business correspondence, short report or proposal, based on a rubric and input text(s); 120–140 words.	One compulsory task	Correspondence: e.g. explaining, apologising, reassuring, complaining. Report: e.g. describing, summarising. Proposal: e.g. describing, summarising, recommending, persuading.
LISTENING 40 minutes	**Part 1**	A gap-filling task involving three short monologues or dialogues of approximately 1 minute each. Each extract is heard twice.	12	Listening for note-taking.
	Part 2	A multiple-matching task involving two sections of five short monologues.	10	Listening to identify topic, context, function, etc.
	Part 3	A multiple-choice task involving a monologue, interview or discussion lasting approximately 4 minutes, heard twice.	8	Following the main points and retrieving specific information from the text.
SPEAKING 14 minutes	**Part 1**	A conversation between the interlocutor and each test taker (spoken QUESTIONS).	Various	Giving personal information. Talking about present circumstances, past experiences and future plans, expressing opinions, speculating, etc.
	Part 2	A 'mini-presentation' by each test taker. The test takers are given a choice of three business-related topics and have 1 minute to prepare a piece of extended speech lasting approximately 1 minute.	One presentation per test taker	Organising a larger unit of discourse. Giving information and expressing and justifying opinions.
	Part 3	A collaborative task. The test takers are presented with a discussion on a business-related topic and the interlocutor extends the discussion with prompts on related topics.	Various	Initiating and responding, negotiating, collaborating, exchanging information, expressing and justifying opinions, agreeing and/or disagreeing, suggesting, speculating, comparing and contrasting, decision-making.

Reading example

General description	
PAPER FORMAT	The paper consists of a range of business-related texts and accompanying tasks. A text may consist of several short sections.
TIMING	One hour
NO. OF PARTS	There are five parts. Parts 1 to 3 test takers' reading comprehension. Parts 4 and 5 test takers' understanding of written English at word, phrase, sentence and paragraph level.
NO. OF QUESTIONS	45
TASK TYPES	Matching 4-option multiple choice 4-option multiple-choice cloze Proofreading.
TEXT TYPES	Informational texts, articles and reports.
LENGTH OF TEXTS	Approximately 150–550 words per text.
ANSWER FORMAT	Test takers indicate their answers by shading a box or writing a word on a machine-readable answer sheet.
MARKS	All questions carry one mark.

Appendix IV – Advice for item writers

Advice on choosing texts

The definition of 'text' in this Manual follows that given in Section 4.6 of the CEFR. It refers to any piece of language, whether spoken or written.

Item writers should be given guidance on selecting texts. It should cover the following points:

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

Reasons for rejecting texts can also be given. They can include:

- a great assumption of cultural or local knowledge (unless this is being specifically tested)
- topics which may be seen as unsuitable for the target test takers. These might include war, death, politics and religious beliefs, or other topics which may offend or distress some test takers
- topics outside the experience of test takers at the target age group
- too high or low a level of difficulty of vocabulary or concepts
- technical or stylistic faults or idiosyncrasies
- poor editing of the original text.

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

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

Advice on presentation

Item writers can be guided on 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 task

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

- An authentic text, around 200 words long, is required. It should have a short title. The emphasis is on single structural words. There should not be a heavy load of unfamiliar vocabulary.
- There should be a minimum of 16 items, more if possible, to allow for selection after pretesting. The first item will be used as an example, and may be numbered '0' (zero). Items should test prepositions,

pronouns, modifiers, verb auxiliaries, etc. They should be spread evenly through the text, and care should be taken that failing to get one right does not lead automatically to also getting the following one wrong (interdependency of items).

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

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

Experienced writers of text-based items often gather suitable texts on an ongoing basis from the recommended sources. When they are asked for items, they select and work on the most promising texts from those already collected. For writing some types of items (e.g. items focusing on grammar or vocabulary), it is useful for the item writer to have a dictionary and thesaurus to hand. When writing listening materials, it is helpful to listen to the passage, so that the test items can be developed directly from the spoken text rather than from the written text on the page.

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

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

Text:
Is the text topic accessible/culturally acceptable/etc.?
Is the text free of any potentially insensitive content?
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 16 items?
Has a suitable title been included?
Items:
Has the required number of items been generated?
Are the items spread evenly through the text?
Is a good range of language focused on?
Has a check been made that all items are structurally focused?
Is it certain that there are no interdependent items?
Have one or two extra items been included?
Have idiosyncratic items been avoided?
Rubric and key:
Has the rubric been checked?
Has an example been provided?
Has a comprehensive key been provided on a separate sheet?

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

Appendix V – Case study – editing an A2 task

This appendix shows the changes made to a task during an editing process and the reasons for these changes. Each new version is displayed with comments on changes below it. Parts of the text which are discussed are highlighted in red.

Version 1 – Submitted by the item writer for editing (first meeting)

Complete the conversation between two friends.

What does Josh say to his friend Marta?

For questions **1–5**, mark the correct letter **A–H** on your answer sheet.

Example

Marta: Hello, Josh. It's good to see you. How was your holiday?

Josh: **0** ..... **E**

Answer: 0 | A B C D **E** F G H

Marta: Where did you go this year?

Josh: **1** .....

Marta: What was the weather like?

Josh: **2** .....

Marta: Great! Did you take any photos?

Josh: **3** .....

Marta: Did you stay in a hotel?

Josh: **4** .....

Marta: It sounds like you had a great time. Will you go again?

Josh: **5** .....

Marta: It would be more interesting to do that.

A	It was quite warm. I went swimming in the sea.
B	No we didn't. My uncle has some friends there and we stayed with them.
C	I thought they were really great.
D	Probably not. I'm planning to go somewhere else next year.
E	It was great, thanks.
F	I took some really good ones.
G	We didn't have enough money.
H	I went with my uncle to Iceland.

Key: 1 H, 2 A, 3 F, 4 B, 5 D

Review of the version submitted for editing (first meeting)

At the (first) editing meeting the writer was asked to resubmit the task with the following changes:

- avoid the consistent answer/question pattern throughout the conversation
- avoid the repetition of certain lexis
- amend the distractors **G** and **C** and related text
- rephrase the lexis and structures outside the vocabulary list and grammatical SPECIFICATIONS.

The first change was necessary to avoid the task being too easy and focusing on discrete answers and questions. In the original version, each gap tested Josh's response to a question put by Marta. The writer was asked to vary the interaction pattern (for example, by turning some of options **A**–**H** into questions) and to rephrase parts (for example, by adding an offer to the end of option **F**) in order to create a more cohesive dialogue.

The second change was to avoid the same verb form appearing in both the question and the answer and making the task too easy. For example, 'Did you take any photos?' and 'I took some really good ones'; 'Where did you go this year?' and 'I went with my uncle to Iceland.' In addition, the writer was asked to vary the lexis. For example, 'great' appeared four times.

The third change was to avoid the distractors **C** and **G** being possible as keys for some items. The writer was asked to rephrase **C** and **G**, and related text, so that they could not fit item 3 and also to ensure that **G** was not too tempting for item 4.

The fourth change was related to the difficulty level of task contents. For example, the writer was asked to rephrase 'probably' which is not in the vocabulary list and 'It would be more …', which was not in the list of functions for this test.

Version 2 – The rewritten task that was re-submitted by the writer

Changes made by the item writer:

i. the interaction pattern is more varied; 'probably' and 'it would be' have been removed.

ii. option C has been changed so it no longer works for item 3.

iii. the text before and after the gap for item 4 has been amended to rule out G for items 3 and 4.

iv. the verb 'took' has been removed from option F.

v. the item writer argued that replacing 'went' with, for instance, 'visited' would make the dialogue unnatural so the two forms of 'to be' have been disguised in longer segments of text.

Complete the conversation between two friends.

What does Josh say to his friend Marta?

For questions **1–5**, mark the correct letter **A–H** on your answer sheet.

Example:

Marta:	Hello, Josh. It's good to see you. How was your holiday?
Josh:	**0** ..... **E**

Answer: 0 — A B C D **E** F G H

Marta:	Where did you go this year? To your uncle's again?
Josh:	**1** .....
Marta:	No. It's too cold for me there.
Josh:	**2** .....
Marta:	Great! Did you take any photos?
Josh:	**3** .....
Marta:	Yes, please! Did you stay in a hotel?
Josh:	**4** .....
Marta:	That was lucky!
Josh:	**5** .....
Marta:	I didn't know that. You must tell me more about it soon.

A	No, in summer it's quite warm. You can even swim in the sea.
B	My uncle has some friends there and we stayed with them.
C	No, but did you enjoy your holiday?
D	Yes, it was! Hotels are very expensive there.
E	It was great, thanks.
F	Lots. I'll bring them to show you if you like.
G	We didn't have enough money.
H	We did something different – we went to Iceland. Have you been there?

Key: 1 H, 2 A, 3 F, 4 B, 5 D

Version 2 – The rewritten task that was re-submitted by the writer after discussion in editing (second meeting)

Complete the conversation between two friends.

What does Josh say to his friend Marta?

For questions **1–5**, mark the correct letter **A–H** on your answer sheet.

Example

Marta: Hello, Josh. It's good to see you. How was your holiday?

Josh: **0** ..... **E**

Answer:

0	A	B	C	D	E	F	G	H
	☐	☐	☐	☐	■	☐	☐	☐

Marta:	Where did you go this year? To your uncle's again?
Josh:	**1** .....
Marta:	No. It's too cold for me there.
Josh:	**2** .....
Marta:	Great! Did you take any photos?
Josh:	**3** .....
Marta:	Yes, please! Did you stay in a hotel?
Josh:	**4** .....
Marta:	That was lucky!
Josh:	**5** .....
Marta:	I didn't know that. You must tell me more about it soon.

A	No, in summer it's quite warm. You can even swim in the sea.
B	My uncle has some friends there and we stayed with them.
C	No, but did you enjoy your holiday?
D	Yes, it was! Hotels are very expensive there.
E	It was great, thanks.
F	Lots. I'll bring them to show you if you like.
G	We didn't have enough money.
H	We did something different – we went to Iceland. Have you been there?

Key: 1 H, 2 A, 3 F, 4 B, 5 D

Review of the version re-submitted for editing (second meeting)

At the (second) editing meeting for this task the following changes were made:

- 'his friend' was deleted from the second line of the RUBRIC.
- Marta's second turn (between gaps 1 and 2) was amended (to 'No. Isn't it very cold there?').
- option **A** was amended to 'Not really. In the summer you can even swim in the sea'.
- option **B** was amended to 'We have some friends there and we slept at their house'.

The first change was stylistic to avoid the repetition of 'friend', which appears in the first line of the rubric and to standardise the rubric.

There were two reasons for the change to Marta's second turn. The first was to rule out option **B** from item 2 (the key is **A**). The second was to provide a stronger indication of the content of the gap. In the initial version, it is possible for the conversation to change in focus after Marta's turn, but by amending this to a question, more support was given for Josh's response.

Option **A** was then changed due to Marta's second turn becoming a question. 'Not really' became the response to Marta's question and the information about the summer and swimming in the sea was retained but amended slightly.

There were also two reasons for the changes to option **B**, which is the key to item 4. The first was to delete the reference to Josh's uncle because this appeared to root the option at the beginning of the conversation rather than in the fourth gap. The reference 'them' in **B** was also potentially confusing. It was decided to make a distinction between the house belonging to his uncle's family and a house that belonged to friends of Josh's uncle. If test takers had missed this distinction, they might not have considered option **B** for item 4. Option **B** was therefore amended to, 'We have some friends there'. The second reason for changing option **B** was to avoid a match at a purely lexical level. 'Stay' appears in Marta's question before item 4 and it also appears in option **B**. 'We stayed with them' was therefore changed to 'We slept at their house'.

Version 3 – The version to be pretested – incorporating changes made at the second editing meeting

Complete the conversation between two friends.

What does Josh say to Marta?

For questions **1–5**, mark the correct letter **A–H** on your answer sheet.

EXAMPLE:

Marta: Hello, Josh. It's good to see you. How was your holiday?

Josh: **0** ..... **E**

Answer:

0	A	B	C	D	E	F	G	H
	☐	☐	☐	☐	■	☐	☐	☐

Marta: Where did you go this year? To your uncle's again?

Josh: **1** .....

Marta: No. Isn't it very cold there?

Josh: **2** .....

Marta: Great! Did you take any photos?

Josh: **3** .....

Marta: Yes, please! Did you stay in a hotel?

Josh: **4** .....

Marta: That was lucky!

Josh: **5** .....

Marta: I didn't know that. You must tell me more about it soon.

A	Not really. In the summer you can even swim in the sea.
B	We have some friends there and we slept at their house.
C	No, but did you enjoy your holiday?
D	Yes, it was! Hotels are very expensive there.
E	It was great, thanks.
F	Lots. I'll bring them to show you if you like.
G	We didn't have enough money.
H	We did something different – we went to Iceland. Have you been there?

Key: 1 H, 2 A, 3 F, 4 B, 5 D

Review of the pretested version (third meeting)

It was decided at this meeting that no changes to the task were necessary. The statistics indicated that the task was at the right difficulty level (see Appendix VII for a description of how to understand these statistics). The target mean difficulty for KET is -2.09 and this task had a mean difficulty of -2.31. Items 1–5 fell within the acceptable RANGE of difficulty, which is -3.19 to -0.99.

	Item Difficulty (LOGITS)
1	-2.72
2	-2.90
3	-2.86
4	-1.92
5	-1.13
MEAN	-2.31

Ruling out possible double keys was also a consideration at this meeting. The breakdown of test taker answers is shown in the Classical statistics report below. For example, for item 2, 20% of the low group chose **F** and for item 4, 50% of the low group chose **D**. These options were re-checked to see if they could be possible answers for items 2 and 4. It was decided that they could not. For example, **F** cannot be the key for item 2 because there is nothing for 'I'll bring them to show you' to refer to. **D** is ruled out as the key for item 4 by 'Yes it was!'.

Item Statistics					Alternative Statistics				
Seq No.	Scale -Item	Prop. Correct	Disc. Index	Point Biser.	Alt.	Prop. Total	Endorsing Low	Endorsing High	Point Biser.	Key
1	1-13	.73	.41	.40	A	.07	.15	.00	-.24	
					B	.07	.15	.00	-.27	
					C	.09	.11	.02	-.13	
					D	.01	.00	.00	-.04	
					E	.01	.00	.00	-.06	
					F	.01	.00	.00	.05	
					G	.02	.02	.00	-.04	
					H	.73	.57	.98	.40	*
					Other	.00	.00	.00		
2	1-14	.76	.41	.36	A	.76	.57	.98	.36	*
					B	.04	.07	.00	-.08	
					C	.05	.07	.00	-.11	
					D	.03	.04	.00	-.06	
					E	.03	.04	.00	-.15	
					F	.07	.20	.02	-.30	
					G	.00	.00	.00		
					H	.02	.02	.00	-.04	
					Other	.00	.00	.00		
3	1-15	.76	.41	.38	A	.02	.00	.00	.00	
					B	.03	.00	.00	-.02	
					C	.07	.15	.04	-.15	
					D	.03	.09	.02	-.16	
					E	.04	.11	.00	-.21	
					F	.76	.50	.91	.38	*
					G	.04	.11	.02	-.18	
					H	.01	.02	.00	-.15	
					Other	.01	.00	.00	-.08	
4	1-16	.58	.56	.45	A	.01	.00	.00	.01	
					B	.58	.28	.84	.45	*
					C	.02	.00	.04	.12	
					D	.29	.50	.07	-.37	
					E	.01	.02	.00	-.04	
					F	.00	.00	.00		
					G	.10	.20	.04	-.22	
					H	.00	.00	.00		
					Other	.00	.00	.00		
5	1-17	.41	.60	.50	A	.02	.02	.00	-.01	
					B	.07	.09	.07	-.06	
					C	.10	.07	.07	-.05	
					D	.41	.13	.73	.50	*
					E	.17	.35	.02	-.33	
					F	.06	.09	.07	-.04	
					G	.07	.11	.04	-.06	
					H	.09	.15	.00	-.22	
					Other	.00	.00	.00		

Version 4 – The live test version
(same as version 3 – no changes were made)

Appendix VI – Collecting pretest/trialling information

This appendix contains possible questions to ask after PRETESTING or TRIALLING (see Section 3.4.2).

Pretest/trialling feedback from invigilators – all components

Please make comments on the following:

1. **Content:** Range and types of questions/texts/tasks, etc.
2. **Level:** Difficulty, e.g. linguistic/cognitive of the various sections/tasks.
3. **Listening pretests only:** Clarity/speed of delivery, accent of speakers, etc.
4. **Candidature:** What is the approximate age of the students who took the pretest?
5. **Further comments?**

Pretest/trialling feedback from test takers – reading test

1. **Did you have enough time to complete the task? (If 'No', how much longer did you need?)**
2. **Was there any vocabulary that you did not understand in the task?**
   (Please note particular words/expressions which were a problem)
3. **Could you follow the ideas and 'line of argument' of the writer?**
   EASILY/WITH SOME DIFFICULTY/WITH GREAT DIFFICULTY
4. **How familiar were you with the subject matter of the passage?**
   VERY FAMILIAR/QUITE FAMILIAR/NOT VERY FAMILIAR/UNFAMILIAR
5. **When (if at all) do you plan to take the live test?**
6. **Are there any other comments you would like to make?**

Pretest/trialling feedback from raters – writing test

About the task **INPUT**: Focus

1. Was the task generally understood?
2. Was the role of the writer clearly identified?
3. Was the target reader clearly identified?
4. Is there any cultural bias? Does the task favour a test taker of a particular background or age?
5. Is any rewording of the question required? If so, please outline your suggestions.

About the task **INPUT**: Language

6. Was the question understandable by B2 level test takers?
7. Was there any confusion over or misinterpretation of the wording?
8. Were test takers confused about the appropriate REGISTER to use?
9. Is any rewording of the question required? If so, please outline your suggestions.

About the task **OUTPUT**: Content

10. Was the task type interpreted appropriately?
11. Were any content points misinterpreted/omitted? Please give details.
12. Was the word length appropriate for the task?

About the task **OUTPUT**: Range/Tone

13. Was any language lifted from the question? Please specify.
14. Which register (formal, informal, etc.) did the test takers use?

About the task **OUTPUT**: Level

15. Did the question give C1 level test takers enough scope to show their ability?

Mark scheme:

16. Please outline suggestions for amendments to the mark scheme.

Overall impression:

17. Please give your overall impression of the question.

Appendix VII – Using statistics in the testing cycle

Collecting and analysing test data requires planning and resource, but can greatly enhance the quality of a test and the interpretability of the outcomes. At the very least it is necessary to record information on the test takers who entered and the mark or grade they achieved, and this can be summarised using simple statistics; see, for example, Carr (2008).

More precise data on test taker performance can show how well items worked and where the focus of editing should lie. Simple-to-use software packages are available to do many of the analyses described below and may be used with small numbers of test takers (e.g. 50).

Additional data to collect includes:

- task level data: the mark achieved by a test taker for each task, and not simply the total score
- item RESPONSE data: the response made by a test taker to each item in the test
- demographic information about the test takers: age, gender, first language etc.

The data

Most Classical analysis software will expect data in something like the form shown in Figure 18. You can use any word-processing application to enter data, but take care to:

- use a fixed-width font, like Courier
- not use tabs
- save the file as plain text (.txt).

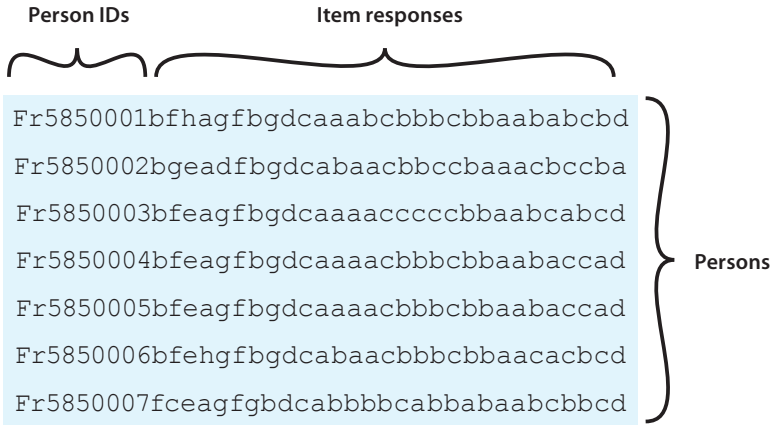


Figure 18 ***A typical file of item response data***

In Figure 18:

- each row has the responses for one person
- the first columns contain an ID for the person (this may include demographic information)
- each column contains the responses to one test item.

This example is for a multiple-choice test where the actual options (a–h) selected by each person are captured.

The analysis software will need some additional information, such as which option is the correct one for each item.

Classical item analysis

Classical ITEM ANALYSIS is used:

- to analyse pretest data and inform the selection and editing of tasks for live use
- to analyse live test response data.

It provides a range of statistics on the performance of items and the test as a whole. In particular:

Statistics describing the performance of each item:

- how easy an item was for a group of test takers
- how well the items discriminate between strong and weak test takers
- how well the key and each distractor performed.

Summary statistics on the test as a whole, or by section, including:

- number of test takers
- mean and STANDARD DEVIATION of scores
- a reliability estimate.

Below we offer some guidelines on what are acceptable values for some of these statistics. They are not intended as absolute rules – in practice the values typically observed depend on the context. Classical statistics tend to look better when you have:

- larger numbers of items in a test
- more test takers taking the test
- a wider range of actual ability in the group taking the test

and conversely, to look worse when you have few items or test takers, or a narrow range of ability.

Figure 19 shows example item statistics based on output from the MicroCAT item analysis package (see Tools for statistical analysis, below). It shows the analysis for three items.

	Item Statistics			Alternative Statistics					
Seq. No.	Scale -Item	Prop. Correct	Disc. Index	Point Biser.	Alt.	Prop. Total	Endorsing Low	Endorsing High	Point Biser.	Key
1	1-1	.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	
2	1-2	.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	
3	1-3	.93	.19	.39	A	.93	.81	.00	.39	*
					B	.07	.18	.00	-.39	
					Other	.01	.00	.00	-.03	

Figure 19 ***Example item statistics (MicroCAT item analysis package)***

Facility

FACILITY is the proportion of correct responses (*Prop. Correct* in Figure 19). It shows how easy the item was for this group of test takers. The value is between 0 and 1, with a high number indicating an easier item. Figure 19 shows that item 1 is the most difficult, item 3 the easiest.

Facility is the first statistic to look at, because if it is too high or low (e.g. outside the range 0.25–0.80%) then it means that the other statistics are not well estimated – we do not have good information from this group of test takers. If they represent the live test population then we can conclude that the item is simply too easy or difficult. If we are not sure about the level of the test takers then it might be that the item is fine but the group is of the wrong level. A practical conclusion is that we should always try to pretest on test takers of roughly the same level as the live test population.

Discrimination

Good items should distinguish well between weaker and stronger test takers. Classical item analysis may provide two indices of this: the DISCRIMINATION index and the point biserial CORRELATION (*Disc. Index* and *Point Biser.* in Figure 19).

The discrimination index is a simple statistic: the difference between the proportion of correct answers achieved by the highest scoring test takers and that achieved by the lowest (usually the top and bottom third of test takers). The output in Figure 19 shows these in the *Low* and *High* columns. For item 1 the difference between the high and low group is (0.66–0.13). This is the value of the discrimination index (within rounding error).

A highly discriminating item has an index approaching +1, showing that the strongest test takers are getting this item correct while the weakest are getting it wrong.

Where the facility is very high or low then both the low and high groups will score well (or badly), and thus discrimination will be underestimated by this index. Item 3 illustrates this problem: 1.00 – 0.81 = 0.19, a low value.

The point biserial involves a more complex calculation than the discrimination index and is more robust to high or low facility. It is a correlation between the test takers' scores on the item (1 or 0) and on the whole test.

In general, items with a point biserial correlation of greater than 0.30 are considered acceptable. A negative point biserial means that strong test takers are more likely to get the item wrong. In this case, check whether one of the distractors is actually a correct answer, or the key is wrong.

Distractor analysis

Distractors are the incorrect options in a multiple-choice item. We expect more low-ability test takers to select a distractor, while high-ability test takers will select the key (correct option, marked '*').

A distractor analysis shows the proportion of test takers choosing each distractor (*Prop.Total* in Figure 19). Item 1 in Figure 19 has quite low facility for the correct option B: 0.38. Distractor D attracts more responses (facility = 0.49). Distractor A actually attracts no test takers at all, so it is clearly not a good distractor. But overall the item works well, with good discrimination, so there is no need to change it. In practice it is difficult to find three distractors that all work well.

The analysis in Figure 19 also shows the low–high proportion choosing, and the point biserial for each option. A good item will have positive point biserial for the key and a negative one for each distractor.

Reliability of scores

There are several ways in which to estimate reliability and several formulae to do so. Each method has different assumptions. The split half method involves dividing the test into two equal parts and comparing the same test taker's score on both parts. With this method, it is important the two halves are as equivalent as possible: they cover equivalent areas of the construct, they are of equivalent difficulty, etc.

Other methods involve measuring the internal consistency of the test. This works well where items are very similar in type and content. However, where items are heterogeneous, the reliability will be underestimated.

For Classical analysis:

minimum number of test takers: 50 to 80 (Jones, Smith and Talley 2006:495)

more information: Verhelst (2004a, b); Bachman (2004)

Rasch analysis

RASCH ANALYSIS is the simplest and most practical form of ITEM RESPONSE THEORY or 'IRT'. It provides a better understanding of item difficulty than Classical analysis and has more additional applications, such as linking test forms.

With Rasch analysis:

- the precise difference in difficulty between two items is clear, as items are placed on an INTERVAL SCALE measured in logits
- the difference between items and test takers, test scores or cut-off points can be understood in the same way, as all these things are measured on a single scale
- item difficulty can be understood independent of the influence of test taker ability (with Classical analysis, a group of able test takers may make an item look easy, or a weak group may make it look hard).

These characteristics mean that Rasch analysis is useful to monitor and maintain standards from session to session. However, to use Rasch analysis in this way, items in different tests must be linked. For example, two tests may be linked in a number of ways:

- some common items are used in both tests
- a group of ANCHOR ITEMS are administered with both tests
- some or all items are CALIBRATED before being used in a live test (see Section 3.4.2, pretesting)
- some test takers may take both tests.

When the data from both tests is analysed, the link provides a single frame of reference for all items, test takers, etc., and items receive calibrated difficulty values. Other tests can be added to this frame of reference in the same way.

Standards can be monitored by comparing the relative position of important elements:

- are the items at the same level of difficulty in all test forms?
- are the test takers of the same ability?
- do the cut-off points (measured in logits) coincide with the same RAW SCORE (also now measured in logits) in all test forms?

Standards can be maintained if the cut-off points are set at the same difficulty values each time.

It is easier still to maintain standards and test quality if tests are constructed using calibrated items. The overall difficulty of a test can be described with its mean difficulty and the RANGE. Test difficulty can be controlled by selecting a group of items which fit the target range and match the target mean.

When you begin to calibrate items, the difficulty values will not mean much. But over time you can study the real-world ability of test takers to give meanings to points on the ability scale. Alternatively, or at the same time, you can make subjective judgements about the items ('I think a threshold B1 learner would have a 60% chance

of getting this item right') in order to give meaning to the item difficulties. In this way the numbers on the ability scale become familiar and meaningful.

For Rasch analysis:

minimum number of test takers: 50 to 80 (Jones, Smith and Talley 2006:495)

more information: Verhelst (2004d); Bond and Fox (2007)

Statistics for marking and rating

Clerical marking

It is important to decide whether markers are performing well. If they are not, action – such as retraining – can be taken (see Section 5.1). If the number of test takers is small, it may be possible to verify the mark given to each item by each marker. However, for larger numbers of test takers, a sample (perhaps 10%) of their work can be reviewed and an error rate established. An error rate is the number of errors a marker has made divided by the number of items they have marked. If this sample is representative of all of their work, the error rate is likely to be the same for all of their work.

In order for the sample to be representative of a marker's work, it is usually best to collect it randomly. To ensure a truly random sample, it is important to consider the way the marker has been working. A random sample does not mean any 10% of papers, otherwise the sample may include only recent work because it is more accessible. In this case, the error rate may be underestimated for the whole time they have been marking: the marker has probably improved during their time doing the job.

Rating

The performance of RATERS can be assessed statistically in a very simple way by calculating the mean of their ratings and the STANDARD DEVIATION (a measure of the spread of their ratings, from low to high). Different raters can be compared to each other and any who stand out as different from the others can be investigated further. This will work if the exam material is given to raters randomly. If it is not, one rater may be asked to rate test takers who are normally better or worse than the average. In this case, the mean will be higher or lower than the other raters but the performance of this rater may be good.

If some tasks are marked by two raters, the reliability of these marks can be estimated. This can be done, for example, in Excel using the 'Pearson' correlation function. The data can be set out as follows:

	rater 1	rater 2
test taker 1	5	4
test taker 2	3	4
test taker 3	4	5
…	…	…

The correlation coefficient will be between -1 and 1. In most circumstances, any number lower than 0.8 should be investigated further, as it suggests that the raters did not perform similarly.

A reliability estimate, like the Alpha produced by MicroCAT (see Tools for statistical analysis, below), can be calculated for the entire group of raters. Data can be laid out as described in Figure 18, with a few modifications: each row can be used to represent the performance of one test taker on one task; the columns can be used for raters' marks.

Many-Facet Rasch Measurement (MFRM)

A more sophisticated way to judge rating performance is to use a technique like MANY-FACET RASCH MEASUREMENT (MFRM). It is a variant of Rasch analysis. MFRM can be done using software called Facets (Linacre 2009). The analysis measures the difficulty of tasks and the ability of test takers as with Rasch analysis, but can also assess the severity and leniency of raters. In addition, it can provide fairer scores for test takers, as severity and leniency effects can be removed.

An important consideration when using MFRM is to ensure that the data contains links between raters, between test takers, between tasks and other facets measured. For example, some performances must be rated by more than one rater to provide a link between raters. Some test takers must complete more than one task to provide a link between tasks. If separated pockets of data are formed, MFRM may not be able to provide estimates for all elements.

For MFRM:

minimum number of performances: 30 for each task to be rated (Linacre 2009)

minimum number of ratings by each rater: 30 (Linacre 2009)

more information: Eckes (2009)

Construct validation

Verifying test structure

Factor analysis or Structural Equation Modelling can help to verify whether the items of the test cover the construct which is intended. The test should display a pattern that can be identified with the model of language use adopted (see Section 1.1). Factor analysis is very useful in the stages of test development, as it is most commonly used to verify that the test, or SPECIFICATION, works as expected.

For factor analysis:

minimum number of test takers: 200 (Jones, Smith and Talley 2006:495)

more information: Verhelst (2004c)

Detecting item bias

Item bias occurs when items unfairly favour or disfavour certain groups of test takers of the same ability. For example, an item may be easier for female test takers than male test takers, even though they are of equal ability. This is unfair because the aim of the test is to measure differences in language ability and not in gender (see Section 1.4).

Care should be taken when diagnosing bias, however, as not all differences between groups are unfair. Learners with a particular L1 may find an item more difficult than learners of the same ability in another group due to differences between the mother tongue and the target language. In the context of measuring language proficiency, this must be accepted as part of the nature of proficiency in the target language, not a problem in measuring it.

One approach to minimising bias is to use a Differential Item Functioning (DIF) methodology to detect possible bias so that it may be investigated further. This involves comparing the responses of groups of test takers who are equally able. For example, if the test is intended to be for adults of all ages, the performance of younger and older adults with approximately the same ability (according to the test) can be compared. Analysis based on IRT is well suited to do this.

For DIF analysis with Rasch analysis:

minimum number of test takers: 500, with at least 100 per group (Jones, Smith and Talley 2006:495)

more information: Camilli and Shepard (1994); Clauser and Mazor (1998)

Verifying test taker sample

When any kind of analysis or research is done using test data, the data should normally be representative of the target group of test takers (the population). Information about these test takers can be collected regularly and checked to see if analysis is being done on a fully representative sample of test takers.

Data on background characteristics of the test takers can be collected whenever a test form is administered (see Section 4). Characteristics can be compared using simple percentages, e.g. comparing the balance of males with females in two different samples.

A more sophisticated analysis will also try to establish whether any differences between two samples are likely to be due to chance. A Chi-square test may be used in this way. The results from an analysis must then be checked qualitatively to see if any differences are likely to cause substantive differences in test taker performance.

Tools for statistical analysis

A number of commercial software packages are available for this kind of work. Some of the calculations can be performed quite easily using Microsoft Excel, or other common spreadsheet programs. Specialist providers are listed in alphabetical order and may provide tools for various types of analysis. Student or demo versions are sometimes available.

Assessment Systems	http://www.assess.com/softwarebooks.php
Curtin University of Technology	http://lertap.curtin.edu.au/index.htm
RUMM Laboratory	http://www.rummlab.com.au/
Winsteps	http://www.winsteps.com/index.htm

Some other free tools are available for specific purposes:

William Bonk, University of Colorado	http://psych.colorado.edu/~bonk/
Del Siegle, University of Connecticut	http://www.gifted.uconn.edu/siegle/research/Instrument%20Reliability%20and%20Validity/Reliability/reliabilitycalculator2.xls

Appendix VIII – Glossary

action-oriented approach
A way to think about language ability where language is seen as a tool to perform communicative 'actions' in a social context.

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

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

assessment use argument
The part of a validity argument which justifies how test results should be interpreted for a particular use.

authenticity
The degree to which test tasks resemble real-life activities. For example, taking lecture notes in a test of academic language competence, rather than just listening. Also see *test usefulness*.

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

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

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

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

construct
A hypothesised ability or mental TRAIT which cannot necessarily be directly observed or measured, for example, in language testing, listening ability.

correlation
The relationship between two or more measures, with regard to the extent to which they tend to vary in the same way. If, for example, test takers tend to achieve similar ranking on two different tests, there is a positive correlation between the two sets of scores.

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

dichotomous item
An item which is scored right or wrong. Dichotomous items include, for example, multiple-choice, true/false and short-answer items.

discrete item
A self-contained item. It is not linked to other items or any supplementary material.

discrimination
The power of an item to discriminate between weaker and stronger test takers. Various indices of discrimination are used. See Appendix VII for details.

domain of language use
Broad areas of social life, such as education or personal, which can be defined for the purposes of selecting content and skills focuses for examinations.

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

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

(standard) error (of measurement)
An estimate of the imprecision of a measurement. For example, if the error of measurement is 2, a candidate with a score of 15 will have a score between 13 and 17 (with 68% certainty). A smaller error will lead to a more precise score.

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

grade
A test score may be reported to the test taker as a grade, for example on a scale of A to E, where A is the highest grade available, B is a good pass, C a pass and D and E are failing grades.

grading
The process of converting test scores into grades.

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

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

interactivity
The degree to which items and tasks engage mental processes and strategies which would accompany real-life tasks. Also see *test usefulness*.

interpretive argument
see *assessment use argument*

interval scale
A scale of measurement on which the distance between any two adjacent units of measurement is the same, but in which there is no absolute zero point.

invigilator
Person who has responsibility to oversee the administration of an exam in the exam room.

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

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

item banking
An approach to the management of test items which entails storing information about items so that tests of known content and difficulty can be constructed.

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

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

linking
Linking is a procedure which 'translates' results from one test or test form so that they can be understood in relation to results of another test or test form. This procedure helps to compensate for differences in test difficulty, or in the ability of test takers.

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

logit
A logit is the unit of measurement used in IRT/Rasch analysis and MFRM.

many-facet Rasch measurement (MFRM)
MFRM is an extension of the basic Rasch model. Item difficulty or test taker ability is broken down into facets so that data relating to these facets can be used to explain the scores given to test takers. For example, rater severity may help to explain scores test takers are given on writing tasks. In this case, scores are seen as being caused by the ability of the test taker, the difficulty of the task and the severity of the rater. It is then possible to remove the influence of rater severity from the final scores given to test takers.

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

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

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

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

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

measurement scale
A measurement scale is a scale of numbers which can be used to measure the difference between test takers, items, cut-off points, etc. on the construct of the test. A measurement scale is constructed by applying statistical techniques to test taker responses to items (see Appendix VII). Measurement scales provide more information than raw scores because they not only show, for example, which test takers are better than others, they also show by how much one test taker is better than another. Nominal and ordinal scales are sometimes referred to as measurement scales but that definition has not been adopted in this Manual.

model fit
When a model (like the Rasch model) is used in statistical analysis, it is important to consider how well the data and model fit each other. A model represents an ideal of how the data should be, so perfect fit is not expected. However, a high level of misfit may mean that the conclusions about the data are false.

model of language use
A description of the skills and competencies needed for language use, and the way that they relate to each other. A model is a basic component of test design.

objectively marked
Items which can be scored by applying a mark scheme, without the need to bring expert opinion or subjective judgement to the task.

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

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

partial credit item
An item scored so that a response which is neither wholly wrong nor right is rewarded. For example, the scores awarded for a response to an item may be 0, 1, 2 or 3, depending on the level of correctness described in the key.

piloting
Trying out test materials on a very small scale, perhaps by asking colleagues to respond to the items and comment.

practicality
The degree to which it is possible to develop a test to meet requirements with the resources available. Also see *test usefulness*.

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

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

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

range
Range is a simple measure of spread: the difference between the highest number in a group and the lowest.

Rasch analysis
Analysis based on a mathematical model, also known as the simple logistic model, which posits a relationship between the probability of a person completing a task and the difference between the ability of the person and the difficulty of the task. Mathematically equivalent to the one-parameter model in item response theory.

rater
Someone who assigns a score to a test taker's performance in a test, using subjective judgement to do so. Raters are normally qualified in the relevant field, and are required to undergo a process of training and standardisation. In oral testing the roles of raters and interlocutors are sometimes distinguished.

rating scale
A scale consisting of several ranked categories used for making subjective judgements. In language testing, rating scales for assessing performance are typically accompanied by band descriptors which make their interpretation clear.

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

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

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

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

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

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

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

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

stakeholders
People and organisations with an interest in the test. For example, test takers, schools, parents, employers, governments, employees of the test provider.

stakes
The extent to which the outcomes of a test can affect the test takers' futures. Test stakes are usually described as either high or low, with high-stakes tests having most impact.

standard deviation (SD)
Standard deviation is a measure of the spread of scores on a test (or the distribution of other data). If the distribution of scores is normal, 68% of them are within 1SD of the mean, and 95% are within 2SDs. The higher a standard deviation is, the further away from the majority of the data it is.

standard setting
The process of defining cut-off points on a test (e.g. the pass/fail boundary) and thus the meaning of test results.

subjectively marked
Items which must be scored using expert opinion or subjective judgement of the task.

task
What a test taker is asked to do to complete part of a test, but which involves more complexity than responding to a single, discrete item. This usually refers either to a speaking or writing performance or a series of items linked in some way, for example, a reading text with several multiple-choice items, all of which can be responded to by referring to a single rubric.

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

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

test usefulness
Test usefulness (Bachman and Palmer:1996) is the idea that a test is most useful when the balance between validity, reliability, authenticity, interactiveness, impact and practicality is optimal.

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

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

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

validation
The process of establishing the validity of the interpretations of test results recommended by the test provider.

validity
The extent to which interpretations of test results are appropriate, given the purpose of the test.

validity argument
An exhaustive series of propositions and supporting evidence which seeks to substantiate the validity of interpretations recommended for test results.

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

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

This glossary has been compiled from the *Multilingual glossary of language testing terms* produced by the Association of Language Testers in Europe (ALTE Members 1998) and the *Dictionary of Language Testing* (Davies et al 1999), both published by Cambridge University Press in the *Studies in Language Testing* series. Additional entries have been written as required.

Acknowledgements

This Manual is a revised version of an earlier one published by the Council of Europe in 2002 as *Language Examining and Test Development*. That document was a version of one produced by ALTE on behalf of the Council of Europe in 1996 and entitled *Users' Guide for Examiners*.

The Council of Europe would like to acknowledge the contribution made by:
the Association of Language Testers in Europe (ALTE) for undertaking the revision of this document

the editing team for this revision:

David Corkill
Michael Corrigan
Neil Jones
Michael Milanovic
Martin Nuttall
Nick Saville

members of the ALTE CEFR/Manual Special Interest Group and their colleagues for supplying additional material and reviewing draft texts on several occasions:

Elena Archbold-Bacalis
Sharon Ashton
Andrew Balch
Hugh Bateman
Lyan Bekkers
Nick Beresford-Knox
Cris Betts
Margherita Bianchi
Inmaculada Borrego
Jasminka Buljan Culej
Cecilie Carlsen
Lucy Chambers
Denise Clarke
María Cuquejo
Emyr Davies
Desislava Dimitrova
Angela ffrench
Colin Finnerty
Anne Gallagher
Jon-Simon Gartzia
Annie Giannakopoulou
Begona Gonzalez Rei
Giuliana Grego Bolli
Milena Grigorova
Ines Haelbig
Berit Halvorsen
Marita Harmala
Martina Hulešová
Nuria Jornet
Marion Kavallieros
Gabriele Kecker
Kevin Kempe
Wassilios Klein
Mara Kokina
Zsofia Korody
Henk Kuijper
Gad Lim
Juvana Llorian
Karen Lund
Lucia Luyten
Hugh Moss
Tatiana Nesterova
Desmond Nicholson
Gitte Østergaard Nielsen
Irene Papalouca
Szilvia Papp
Francesca Parizzi
Jose Ramón Parrondo
Jose Pascoal
Roberto Perez Elorza
Michaela Perlmann-Balme
Tatiana Perova
Sibylle Plassmann
Laura Puigdomenech
Meilute Ramoniene
Lýdia Ríhová
Shelagh Rixon
Martin Robinson
Lorenzo Rocca
Shalini Roppe
Dittany Rose
Angeliki Salamoura
Lisbeth Salomonsen
Georgio Silfer
Gabriela Šnaidaufová
Ioana Sonea
Annika Spolin
Stefanie Steiner
Michaela Stoffers
Gunlog Sundberg
Lynda Taylor
Julia Todorinova
Rønnaug Katharina Totland
Gerald Tucker
Piet van Avermaet
Mart van der Zanden
Juliet Wilson
Beate Zeidler
Ron Zeronis

Council of Europe reviewers:

Neus Figueras
Brian North
Johanna Panthier
Sauli Takala

the publishing team:

Rachel Rudge
Gary White

The Association of Language Testers in Europe (ALTE), as an International Non-Governmental Organisation (INGO) with consultative status in the Council of Europe, has contributed to the resources which make up the Council of Europe's 'toolkit', including the EAQUALS/ALTE European Language Portfolio (ELP) and the CEFR Grids for Analysis of speaking and writing tasks.

Together with the Council of Europe Language Policy Division, ALTE is keen to encourage users of the toolkit to make effective use of the CEFR in their own contexts to meet their own objectives.

Produced by:
Association of Language Testers in Europe
1 Hills Road,
Cambridge CB1 2EU
United Kingdom
www.alte.org

On behalf of:
The Council of Europe

EMC/6212/1Y04