New revised

Manual for Language Test Development and Examining

For use with the CEFR
Produced by ALTE on behalf of the
Language Policy Programme,
Education Department,
Council of Europe

ALTE Association of Language Testers in Europe

COUNCIL OF EUROPE
CONSEIL DE L'EUROPE

Cover design and layout: ALTE

Cover photos / Inner photos; non-copyrighted pictures

This document was co-produced by the Council of Europe and ALTE.

Manual for Language Test Development and Examining

For use with the CEFR

Produced by ALTE on behalf of the
Language Policy Programme,
Education Department, Council of Europe

Language Policy Programme,
Education Department
Council of Europe (Strasbourg)

www.coe.int/lang

Contents

Foreword

In the 15 years since its publication, the original *Manual for Language Test Development and Examining* (2011), commissioned by the Council of Europe and produced by the Association of Language Testers in Europe (ALTE), has provided an excellent resource for test developers and a key instrument to support the implementation of the Council's flagship instrument, the Common European Framework of Reference for Languages (CEFR), published in 2001.

The same period has been one of unbridled change, not only because of the digital revolution and the arrival of AI but also in relation to the increasingly dynamic and complex linguistic and cultural diversity in our societies. These changes present major challenges to education in general and language education in particular. The Council of Europe's response to these challenges has included, among others, the publication of the CEFR Companion volume (Council of Europe, 2020) with new descriptors for mediation, plurilingual/pluricultural competence and signing competences; and the adoption, two years later, of Recommendation CM/Rec(2022)1 of the Committee of Ministers to member States on the importance of plurilingual and intercultural education for democratic culture. In the same vein, and with a focus on the implications of these changes on language testing, ALTE took the initiative to update and revise the original Manual, a decision warmly welcomed by the Council of Europe and resulting in this joint publication. Together, these instruments broaden the perspective on language education, ensuring a comprehensive and integrated approach to the learning, teaching and assessment of all languages. In so doing they strengthen the original aim of the CEFR 'to promote democratic citizenship, social cohesion and intercultural dialogue.'[1]

On behalf of the Council of Europe, I would like to thank ALTE for its ongoing commitment to these values and am confident that this revised *Manual for Language Test Development and Examining* will prove to be as successful as the original version. I look forward to our continued collaboration.

Sarah Breslin
Executive Director, ECML
Head of Language Policy, Council of Europe

[1] Recommendation CM/Rec(2008)7 of the Committee of Ministers on the use of the Council of Europe's Common European Framework of Reference for Languages (CEFR) and the promotion of plurilingualism, available at https://search.coe.int/cm?i=09000016805d2fb1

Preface

ALTE is pleased to launch this revised *Manual for Language Test Development and Examining* in 2026, and considers it an important contribution to ALTE's ongoing involvement with the Common European Framework of Reference for Languages: Learning, teaching and assessment (CEFR) and to the work of the Council of Europe more generally.

2026 marks the 25th anniversary of the CEFR itself and this Manual had its origins in the years that led up to that publication in 2001. ALTE had collaborated with the Council of Europe Language Policy Division on the development of the Framework itself during the 1990s, and when the pilot version appeared in 1996, ALTE was requested to produce Guidelines for Item Writers for those wanting to develop test materials based on the CEFR and its newly proposed six reference levels. This document was first published in 1997 and later appeared in a revised version published by the Council of Europe, Language Examining and Test Development (2002).

After the CEFR was published in its final form in 2001, several User Guides and other documents were commissioned by the Council of Europe to form part of a toolkit for users with different purposes and uses in mind. One of these was a Manual to help assessment providers align their examinations with the Framework. This was in response to feedback from the language testing community that saw an urgent need for such a manual. Consequently, a 'sounding board' of experts and a smaller Authoring Group were set up to oversee and draft this document. Following a phase of piloting and review, this led to the publication of the Manual for Relating Language Examinations to the CEFR in 2009.

Meanwhile, the Council of Europe agreed with ALTE that it would be a good idea to revise and extend the existing guidelines to produce a second Manual specifically for language test development and examining. The Council therefore commissioned ALTE to carry out this project which then took place over several years leading up to 2011. The first version was jointly published by both organisations, with a Foreword by Joseph Sheils from the Language Policy Division and an Introduction by Dr Michael Milanovic, the Manager of ALTE.

Over the past 15 years, the Manual has been widely used and has had an important impact amongst ALTE members and beyond. However, in light of the changing world of learning, teaching and assessment, especially during and after the Covid pandemic, a review and update of the Manual was thought necessary.

The increasing use of digital technology, and the prospect of using artificial intelligence (AI) in many dimensions of assessment was a major consideration, along with its social and economic impacts. Elsewhere, the 'multilingual turn' in language education, and the greater focus on diversity and inclusion, raised important questions about assessment design and delivery that were no longer adequately covered.

It was therefore decided that a Working Group should be convened to take forward a revision project with the involvement of ALTE members and the approval of the Council of Europe. From the Council's perspective, the publication of the CEFR Companion volume pointed to the need for this revision, and care has been taken by the Working Group to ensure that there is strong alignment between changes made in the revised Manual and the Companion volume. Updated guidance on the alignment of language education to the Framework is available in the Handbook for Aligning Language Education (2022), which ALTE had contributed to (in collaboration with the British Council, EALTA, and UKALTA). This was also taken into account.

The Working Group met regularly over a period of three years and sought to involve the widest possible participation of ALTE members from across all forms of membership. Guidance on technical matters was sought from Expert Members and other specialists, and the final draft was reviewed by independent experts nominated by the Council of Europe. Thanks go to those who have led this project and to the many individuals who have participated with their views and expertise.

Particular gratitude goes to Dina Vîlcu who led the Working Group in the latter phases, and to members of the ALTE Secretariat who provided the necessary coordination.

ALTE is grateful for the support of Sarah Breslin (Executive Director of the ECML and Head of Language Policy at Council of Europe) and appreciate her Foreword to this Manual, which emphasises the importance of the collaboration between ALTE and the Council of Europe in promoting the values of the CEFR.

While this Manual is intended to accompany the CEFR, it is also an up-to-date resource that is freely available to anyone with an interest in language assessment whatever their background. I hope therefore that it will be widely used and look forward to receiving feedback from the widest possible range of users that can inform future revisions.

Nick Saville
Secretary-General, ALTE

Introduction

Background

The first edition of the *Manual for Language Test Development and Examining* (MLTDE) was published in 2011, following the publication of the finalised version of the *Common European Framework of Reference for Languages: Learning, teaching, assessment* (CEFR 2001) by the Council of Europe 10 years earlier, in 2001. During those 10 years, the CEFR has had an overwhelming impact on learning, teaching and assessing languages, meeting the need to facilitate comparisons across language teaching and testing systems in Europe and increasingly beyond, and thus further stimulating the discussion among international experts in the field of language testing.

The original MLTDE was produced by the Association of Language Testers in Europe (ALTE) on behalf of the Council of Europe's Language Policy Division to provide guidelines for the development of language tests from the perspective of the CEFR's action-oriented approach to the conceptualisation of language competence. It thus combined a stocktaking of best practices in language test development in general with a specific view of the underlying language model, which would in turn directly influence construct definition.

Having encouraged the production of a 'toolkit' around the CEFR to expand on specific aspects or facilitate procedures, the Council of Europe then developed several additional language testing resources, including:

- *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR) – A Manual* (Council of Europe, 2009), complemented in 2022 with the handbook on *Aligning Language Education with the CEFR* (British Council et al.)
- 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)
- *Relating Language Examinations to the Common European Framework of Reference for Language: Learning, teaching, assessment* (CEFR) – *Highlights from the Manual* (Council of Europe, ECML, 2011)
- Exemplar materials for illustrating the CEFR levels
- Reference Level Descriptions for several languages: https://www.coe.int/en/web/common-european-framework-reference-languages

ALTE contributed to the resources which make up the toolkit, including the MLTDE, the EAQUALS/ALTE European Language Portfolio (ELP) and the ALTE content analysis grids (partially incorporated into the *Manual for Relating Language Examinations to the CEFR*), encouraging 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 specifically addresses a very complex aspect of validity: the alignment of tests to the CEFR. With its clear focus on alignment, covering all other aspects of test development is beyond its scope. This is the task of the *Manual for Language Test Development and Examining*. It is designed instead to be complementary to the Manual for relating language examinations to the CEFR and the handbook *Aligning Language Education with the CEFR*, as it focuses on aspects of test development and examining not covered in those publications.

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 a common understanding in the fields of language learning, teaching and assessment, and provides a compendious discussion of language education, thereby creating a common language for talking about every aspect of it. It also provides a set of reference levels for identifying levels of language proficiency, ranging from near-beginner (A1) to a very advanced level (C2), and covering a range of different competences and areas of use.

The CEFR Companion volume (Council of Europe, 2020) attempted to provide a more fine-grained definition of the proficiency levels, notably by: adding descriptors to the C scales; including plus level scales; creating the Pre-A1 level; and providing scales for mediation as well as plurilingual and pluricultural competences. It also widened the context of language use described by adding content areas such as online communication and reading for leisure, and by incorporating scales for sign language. Notwithstanding the important implications of these additions for test developers, the overall aim of the CEFR, which is much broader than assessment, remains the same: 'facilitating quality in language education and promoting a Europe of open-minded plurilingual citizens' (Council of Europe, 2020, p.26).

The CEFR and the CEFR Companion volume constitute 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), they cannot be applied to all contexts without some kind of user intervention to adapt them 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' (CEFR 2001, p.iv), 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).

The authors of the second edition of this Manual accordingly set out to suggest, not prescribe, ways in which test developers can make use of the CEFR/CEFR Companion volume's offer, and provide guidelines on how these can be integrated into a useful language testing system.

Whereas 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' (Council of Europe, 2009, p.2), the approach adopted in this Manual focuses on the test development process itself, rather than on specific alignment procedures, and shows how a link to the CEFR can be built into 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 aims to provide a coherent guide to test development in general which will be useful in developing tests for a range of purposes. It 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 1.1 offers an overview of the cycle, which is then elaborated in detail in each chapter.

It is possible to use this Manual if the test provider uses a framework other than the CEFR, as many processes in testing, for instance all procedures related to test delivery, are independent of the test's construct. It is however beyond the scope of this Manual to consider other frameworks in detail.

The readers of this Manual

This Manual is for anyone interested in developing and administering language tests which relate to the CEFR. It is designed to be useful to novice language testers as well as more experienced ones. 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 online, or a single teacher who wishes to test the students in their 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 2001 and the CEFR Companion volume, and will be ready to use them together with this Manual when developing and using tests.

Why a revision?

As language testing is a fertile discipline, the need to review the MLTDE after more than a decade became evident. There are several reasons behind this initiative:

- The Council of Europe was often approached with requests from users to develop the CEFR in different respects: complementing the illustrative scales published in 2001 with descriptors for mediation; reaction to literature and online interaction; producing versions for young learners and for signing competences; and developing more detailed coverage in the descriptors for A1 and C levels (CEFR Companion volume, p.13). These requests were addressed in 2020 through the publication of the CEFR Companion volume; all these new developments needed to be reflected in the MLTDE.
- In a very short time, artificial intelligence (AI) has become a common tool in many testing organizations, which is being used in almost all operations of the Testing Cycle. The new developments in technology, including automated production of test content, digital or hybrid test delivery, remote proctoring, and automated marking and rating of written or spoken test taker performance impact significantly on every step of the Testing Cycle. The users of this Manual need to be made aware of the possible benefits coming with the fast pace of progress and also the caveats they need to consider. In a survey conducted among testing organisations in the early phases of the revision of the MLTDE, technology-related questions and AI were named as the most important topics that should be covered in the second edition. These are topics where information very quickly becomes outdated. We therefore had to find a compromise between topicality and durability, and hope to have included information that will remain relevant, focusing mainly on the process rather than on the product, with questions to ask when considering the use of AI that clearly define the purpose of use and remain aware of possible problems.
- In connection with the point above and the legal developments in the European space it was felt that a focus on data protection issues was necessary.
- Language testing and the outcomes it produces under the form of language certification have become increasingly important at an individual and societal level. With much more intense migration around the world, language testing becomes a means of integration and/or control of access in the new country. Therefore, increased awareness of the societal consequences of testing is needed, in terms of justice, fairness and equity.
- With the new political and social changes in the world, the testing industry became concerned with new groups of test takers, e.g. LESLLA learners (Literacy Education and Second Language Learning for Adults). Special attention needs to be paid to this group of test takers, through adequate data collection on the testing population in the test planning phase and reflection of the findings in the whole testing process.

All these points had to be addressed in the second edition of the MLTDE, although we are fully aware that developments are going on even as this Manual is being published and information, especially in the technical field, can become outdated quickly. We have attempted to provide guidance that will enable practitioners to navigate the constantly changing landscape, rather than focusing narrowly on particular new developments.

Structure and content of this Manual

The structure of the chapters replicates the first edition of this Manual, while the content of each chapter has been thoroughly revised and expanded in order to accommodate recent developments in testing and technology.

While Chapters 1, 2, and 3 are related to the conceptualisation and development of the test, choosing the language model to be used, and designing the construct, specifications, and test content,

Chapters 4 and 5 refer to what in the title of the Manual is called 'examining', which could be called the 'executive' part of the assessment process: the one concerned with the interaction of the test taker with the test and the assessment of the test responses. Chapter 6 concerns all the stages of the Testing Cycle, and ensuring test quality and usefulness.

The chapters

Chapter 1 – *Fundamental considerations* – introduces the fundamental concepts of language proficiency, validity, reliability, fairness and justice.

Chapter 2 – *Developing the test specifications* – goes from the decision to provide a test, a discussion of the test construct, the conception of a provisional test format, item formats and rating criteria, through to the elaboration of final specifications.

Chapter 3 – *Item writing, test construction and content management* – covers item writing and construction of tests, including guidelines on trialling and pretesting.

Chapter 4 – *Delivering tests* – covers the administration of tests, from registration of test takers through to the return of materials. This chapter also addresses questions with regard to online test administration, remote proctoring and data management.

Chapter 5 – *Marking, grading and reporting of results* – addresses the marking process with a view both to human and automated marking and grading, and describes approaches to rater training and monitoring the quality of rating.

Chapter 6 – *Monitoring, review and revision* – describes what data is necessary to monitor the continued functioning of the test, and how this should be implemented over time in order to improve the quality and usefulness of the test.

Each chapter includes key questions for reflection from the readers, and a section of further reading in which relevant titles and digital resources are indicated.

Chapters 2–6 open with the diagram of the *Testing Cycle*, upon which the progress until that point and what is included in that particular chapter are marked.

Chapters 1–5 close with a box summarising the chapter and introducing briefly the next steps in the process.

The appendices

The appendices have been completely reorganised. They contain information which was felt to be important, but which would have interrupted the flow of the chapters. They include examples of procedures and instruments concerning validation, test construction and statistical analysis.

Appendix I – *Building a standards-based validity argument* – exemplifies the construction of a validity argument (a text justifying the use of a given test for a given purpose by presenting evidence on all important aspects of its quality) with the help of the ALTE Minimum Standards. It should be noted that the use of these test quality guidelines is not a prerequisite for test construction in accordance with the principles of the CEFR. Still, the reader may find it useful to follow their structured approach. Even without using the ALTE Minimum Standards, the questions discussed in Appendix I are considered relevant for any kind of test. Each Minimum Standard is accompanied by a description of the practical requirements it calls for.

Appendix II – *Collecting pretest/trialling information* – provides guidance on the collection of relevant information from test takers, administrators and raters at the pretesting/trialling stage.

Appendix III – *Using statistics in the Testing Cycle* – is concerned with the use of statistics in testing, treating concepts which can be analysed quantitatively. It contains sections on describing the test population, minimum sample sizes for representativeness, Classical item analysis, estimating the reliability of test scores, Item Response Theory (IRT) modelling and Rasch analysis, and Differential

Item Functioning (DIF). Basic concepts such as facility, discrimination and distractor analysis are explained in the Classical analysis section, as well as more advanced concepts in the other sections.

Appendix IV – *Glossary* – provides definitions of terms commonly used in this Manual. For a general glossary of testing terms, we refer the reader to the ALTE Multilingual Glossary of Language Testing Terms, accessible at https://www.alte.org/Materials

Information from the appendices of the original version of MLTDE, such as the presentation of the specification development process, has been integrated into the text of the chapters in the present Manual.

How to use this Manual

Even though the principles of language testing introduced here have general applicability, the test providers 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 factors like 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* or in the *Handbook for Aligning Educational Materials to the CEFR*. The reader is also referred to ALTE's 2020 *Principles of Good Practice* for background information on validity and building language tests in general. When talking about concepts that are explained in these publications, this Manual refers the reader to the resources in question.

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. Internal referencing, between chapters, is available, so readers concerned with one particular stage of the Testing Cycle will be directed to relevant sections in other chapters when it is the case. 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 chapter points readers towards resources covering the topics in greater depth, and towards practical tools. This section is followed by key questions, which will help to strengthen understanding of what has been read.

This Manual does not prescribe any specific course of action, but rather 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 Companion volume's illustrative scales.

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.

Test providers have an important role to play in ensuring that the CEFR has a lasting and positive impact: integrating the principles and model of the CEFR into the routine procedures of all stages of the Testing Cycle will enable alignment arguments to be built up and strengthened over time as professional systems develop to support the claims being made and to adapt the descriptors where necessary to suit specific contexts and applications.

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 the MLTDE, or 'this Manual'.

The *Common European Framework of Reference for Languages: Learning, teaching, assessment* (2001) is referred to as CEFR 2001.

The *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume* (2020) is referred to as the CEFR Companion volume.

The CEFR Companion volume as the more recent publication is quoted by default, except when a topic is addressed that is only covered in CEFR 2001.

The conceptual framework as such, comprising both publications, is referred to as CEFR.

Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment – A Manual (Council of Europe, 2009), is referred to as *Manual for Relating Language Examinations to the CEFR*.

The handbook *Aligning Language Education with the CEFR* (British Council et al., 2022) is referred to under its title.

For all other citations the APA convention is used unless otherwise indicated.

1 Fundamental considerations

1.1 What is a test?

While a test may appear to be simply a set of questions, what the test taker sees on exam day is only the surface of the product. In this Manual we use 'test' to mean the larger concept of a tool for measuring something that is defined by specifications and that may have many surface versions. Many processes are required to produce not just this surface, but the evidence that allows us to put trust into its outcome. A vast amount of work goes into ensuring that the test's results are meaningful and trustworthy. This Manual is a guide through that process.

The core framework for this work is the Testing Cycle, which is illustrated in Figure 1 below. It is vital to conceptualise this not as a linear path, but as a circular process. Each stage provides crucial feedback on the test's functioning and its overall fitness for purpose. This information then becomes the basis for reviewing and optimising the test. Note that feedback may require re-thinking and making changes to work that was done in a previous stage and may even influence the way the test's construct is defined or the way the test specifications are set up, so that work on the test versions may feed back on the test development phase. Note also that different test programmes will vary in how much detail goes into each stage, depending on the test use, how complex the construct is, how many test takers there are, and other factors.

It is important to keep in mind that there are many different kinds of language tests. They may differ in their concept of language proficiency, in the scope of what is covered, in their intended use, in the mode of presentation (paper, digital), in the possible consequences of pass or failure, or in the size of the test population, to name but a few. A diagram showing the stages of test development and production in a general way, like the one below, will therefore necessarily have to have a certain level of abstractness. Those working on a given test will need to consider how each of the stages apply

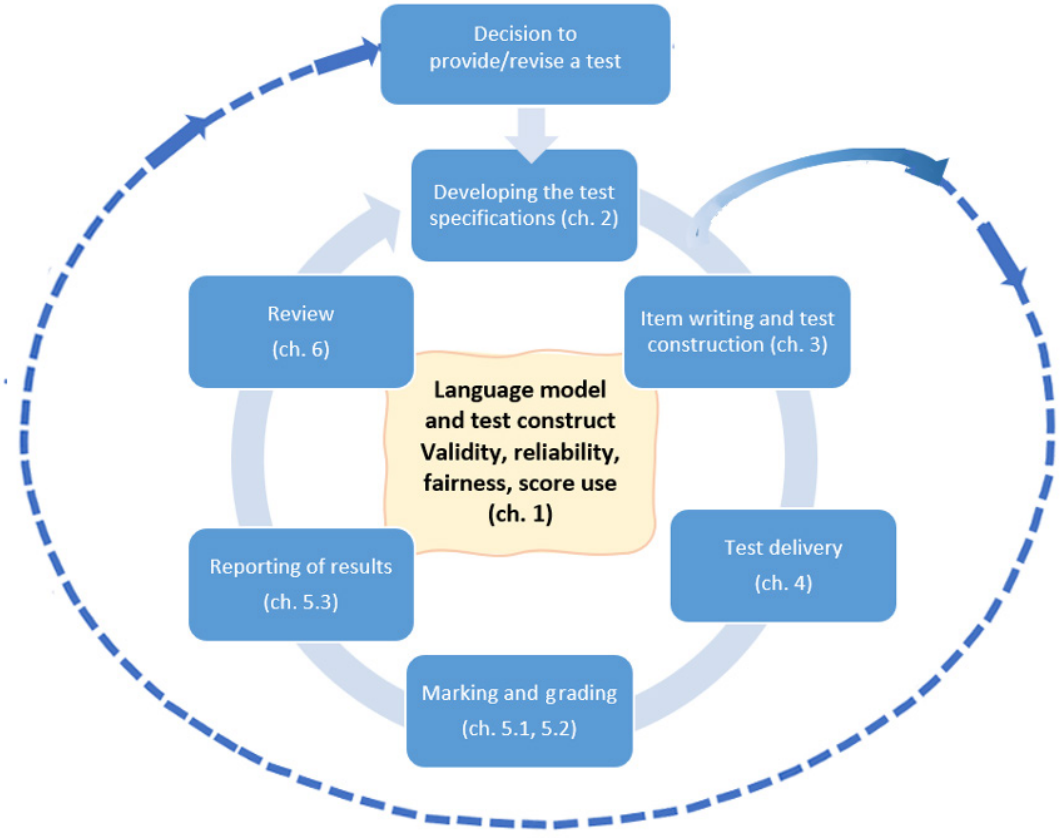


Figure 1 ***The Testing Cycle***

to that particular test. Also bear in mind that the stages of work can be divided in different ways, so that different resources on test development may have different labels and numbers of stages. The general activities are similar, however, across different ways of dividing them up. And all testing programmes should ensure that throughout each stage, testing programmes are referring back to the construct to ensure that the test is functioning as intended (see Sections 1.3.3 here and 2.4.1 in Chapter 2 for discussions of the construct).

This first chapter will present some fundamental considerations that are relevant for any kind of test. The following chapters will look at the stages of the Testing Cycle together with questions that should be considered at each stage to help users find the best way for their own test. Answering these questions will help those responsible for the test to draw up the validity argument that should be the basic document explaining the purpose of the test and showing how its design, implementation, and scoring support that purpose (see Section 1.4 and Appendix I).

1.2 The structure of this chapter

This Manual's practical guidelines for constructing language tests require a sound basis in some principles that underlie all the stages of the Testing Cycle and form the basis for validation. This chapter discusses those principles, and also provides an overview of the overall management of the Testing Cycle.

The chapter is structured as follows:

- **Section 1.3** introduces language proficiency models, the CEFR framework, and the concept of a construct.
- **Section 1.4** explains validation and how to build a validity argument, using the ALTE Minimum Standards framework as an example.
- **Section 1.5** addresses cross-cutting considerations of test impact (including ethics, fairness, and justice), reliability, and addressing practical constraints.
- **Section 1.6** provides guidance on planning and managing the test development process.

1.3 Defining language proficiency

The construct (see Section 1.3.3) is the key concept underlying the Testing Cycle, as noted above: it defines what the test is intended to measure. To develop the construct, it is useful first to identify a model of language use, so before discussing the construct we explore concepts of what it means to 'know' and 'use' language, focusing particularly on the Common European Framework of Reference for Languages (CEFR).

1.3.1 Models of language use and competence

Language in use is a very complex phenomenon which calls on a large number of different competences, such as sociolinguistic, pragmatic and intercultural 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 foundation for deciding which aspects of use or competence can or should be tested (see discussion of the construct here in 1.3.3 and in 2.4.1 in Chapter 2), and helps to ensure that test results will be interpretable and useful.

1.3.2 The CEFR model of language use

1.3.2.1 Descriptive scheme

Influential models of language competence have been proposed by several authors (e.g. Canale and Swain, 1981; Bachman, 1990; Weir, 2005). These works set the scene for the communicative

approach in language assessment and the socio-cognitive framework, which is the foundation for many modern language tests.

For this Manual we focus on the CEFR, as set forth in the original 2001 book and the 2020 Companion volume, which proposes a general model of communicative competence that applies to language use and language learning. The Manual could also be applicable to the development of tests based on frameworks other than the CEFR that share the same approach to language and communication, but a discussion of these approaches is beyond the scope of the Manual.

The CEFR's 'action-oriented approach' is introduced in CEFR 2001 as 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 (p.9; emphasis in original).

This paragraph identifies the major elements of the model, which are then presented in more detail in the text of CEFR 2001 and the CEFR Companion volume. Figure 2 below presents a schematic model of the descriptive scheme; the figure comes from the CEFR Companion volume (p.32), and is based on work by Piccardo et al. (2011). It shows competences sub-divided into two: *General competences* and *Communicative language competences*, each of which is further subdivided. While general competences such as declarative knowledge (savoir) or existential knowledge (savoir-être) are part of a learner's learning endowment, they cannot be the focus of a language test, and are therefore shown greyed out. In addition to competences, proficiency also includes activities and strategies. These are each divided into four modes of communication: reception, production, interaction, and mediation. The CEFR descriptive scales are then categorised according to one of the coloured boxes; the diagram thus captures the broad areas of language and communication that the CEFR model addresses.

The CEFR Companion volume further notes that 'with its communicative language activities and strategies, the CEFR replaces the traditional model of the four skills (listening, speaking, reading, writing), which has increasingly proved inadequate in capturing the complex reality of communication'

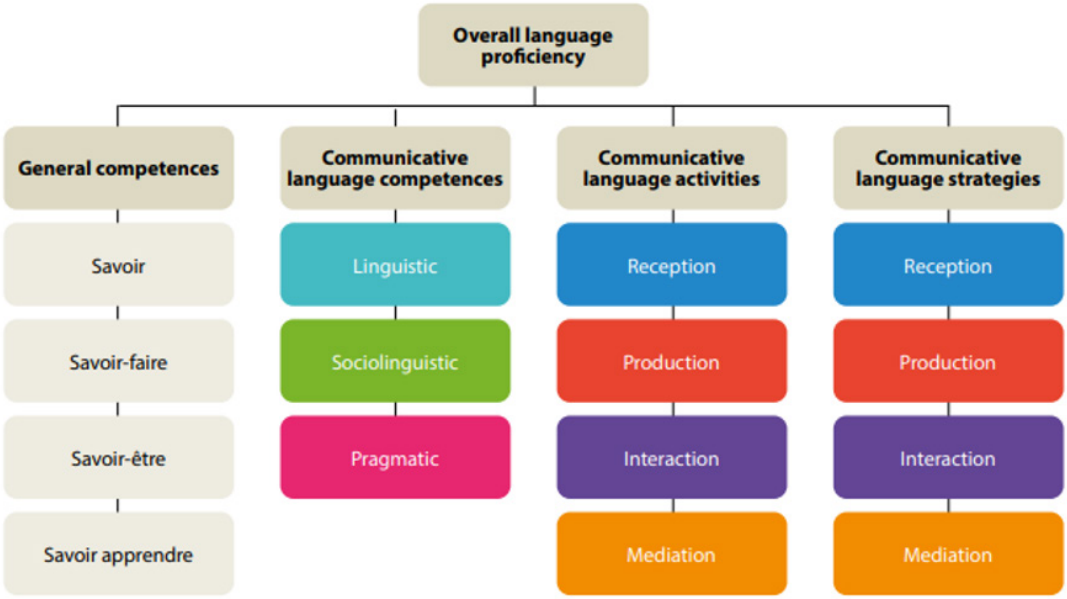


Figure 2 ***The structure of the CEFR descriptive scheme***

(p.33). That being said, many existing tests have sections and report scores according to the traditional four skills labels, and many users of test scores require that scores be reported according to those labels. Language testers therefore need to reflect on how to work with the classifications of the CEFR descriptor scales (see Section 1.3.2.3 below) in the contexts of their own tests and stakeholder needs as they articulate how the tasks and items on their tests reflect the model of language use. In particular, testing organisations should consider the increased importance attached to interaction and mediation in the model articulated in the CEFR Companion volume and the implications of that emphasis for their own test constructs, while at the same time determining which descriptor scales are relevant for interpreting scores according to stakeholder requirements.

1.3.2.2 The integration of competences and modes

The distinct boxes in Figure 2 above may seem to indicate that the competences, modes, activities, and strategies are themselves distinct aspects of the model to be tested independently. However, any communicative act involves many competences and strategies at the same time. For example, when a person 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. Strategies and activities from Reception and Interaction modes are also likely involved in this example. This natural integration creates a challenge for testing, as it is often difficult to determine which competence, strategy, or mode is at play in any given item. See Sections 2.4.2 and 3.1.2.1 for the implications for specifications and task design.

1.3.2.3 The Common reference levels of the CEFR

In addition to the model presented above, the CEFR also identifies a framework of (originally) six levels of communicative language ability (now seven, with the inclusion of 'Pre-A1' in the CEFR Companion volume) 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) overall reading comprehension is:

> 'Can understand very short, simple texts a single phrase at a time, picking up familiar names, words and basic phrases and rereading as required.'

Compare this with a high-level (C2) descriptor for the same scale:

> 'Can understand a wide range of long and complex texts, appreciating subtle distinctions of style and implicit as well as explicit meaning.' (CEFR Companion volume, p.54)

Each scale has descriptors distinguishing the levels with respect to that particular scale. The seven levels are conceptually consistent across scales; for example, A1 is interpreted as 'the lowest level of generative language use,' whereas at A2 the use of language for social functions emerges (CEFR Companion volume, p.173); these general attributes will be instantiated differently for different modes, activities, competences, and strategies, but there is a common understanding that ties these together. The reader is referred to Appendix I of the CEFR Companion volume for a discussion of the interpretation of the levels. See Sections 2.4.2 and 5.2 in this Manual for discussion of the use of the CEFR in content development and rating.

As language testers we should understand the 'Can Do' statements correctly. They are meant to be illustrative. Therefore, they are not exhaustive, prescriptive, a definition, a curriculum, or a checklist. They offer guidance to educators and language testers so that they can recognise and talk about ability levels. They are useful as a guide for test development, but adopting them does not mean that the work of defining ability levels for the test has been completed.

1.3.2.4 User profiles

Given the structure of the framework, with its multiple descriptor scales for different competences, modes, activities, and strategies, and the level descriptors for each scale, a key aspect of the model is that different language users may display different profiles of strengths and weaknesses across different scales. For example, Figure 9 in the CEFR Companion volume shows a hypothetical example of a person with different levels for the scales for overall proficiency for each communicative activity: overall oral production might be at B2, whereas overall written production might be at B1. Even within

a communicative activity, similarly, it is to be expected that a language user who is best described at B1 for overall reading comprehension, for example, may not perform at B1 for every reading-related descriptor scale: the user may have more experience and focus on reading instructions and come close to B2 on that scale, whereas reading for information and argument might not be so important to the user, and the level might be closer to A2. See Section 2.7 of the CEFR Companion volume for a discussion of profiles; the implications for testing are important, as the determination of relevant scales and the expected levels on those scales are an important part of determining the construct. See the discussion of the test construct below and in Chapter 2 (Section 2.4.1).

1.3.3 The construct

Models of language use such as the CEFR are broad foundations for understanding the nature of what is to be tested, but they typically encompass more than any stakeholder wants to test: it would be practically quite difficult for any test to encompass all the descriptor scales, and most stakeholders have particular profiles that they are looking for. The construct is the definition of the aspects of language use that are of importance to the particular target language use (TLU) domain: what will test takers need to be able to do with the language, and in what contexts? Even for a test of general proficiency that covers the broad personal, public, occupational, and educational domains, stakeholders will typically consider some language abilities more important than others. See Chapter 2 (Section 2.4.1) for further discussion of the test construct. Note that the construct is a crucial reference for all stages of the Testing Cycle, not just specifications design: items, tasks, rating schemes, and reporting all need to be evaluated against the construct to ensure valid use of the test scores.

1.4 Validity arguments

1.4.1 Understanding validity

Validity was originally defined as the property of a test to measure what it was intended to measure. This is, however, only a part of the picture: what if the test contains tasks aimed at, say, Level A2 of the CEFR, but is used to place learners into language courses with starting levels between A1 and B2? One of the most important aspects of validity is therefore that it is not the test which is validated, but the inferences about score meaning or interpretation and what consequences these entail (Messick, 1989a; Bachman & Palmer, 1996; Chapelle & Voss, 2021; Kane, 2010). As stated in the ALTE *Principles of Good Practice* (2020), 'validity in language assessment is normally taken to be the extent to which a test can be shown to produce scores which are an accurate reflection of the test taker's true level of language ability … it is important to note that statements about validity should refer to the particular interpretations of test scores for proposed uses, rather than to the test itself – it is the use of the test that needs to be valid in the first instance' (p. 17). This extended definition of validity 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.

Validity is a matter of degree, so we need to evaluate each potential use of the test. For example, if a test of communicative ability in Italian was designed for use in primary or lower secondary schools to test for progress in the A1–A2 range, using scores on that test to screen candidates for a job requiring A2 proficiency would be less valid, and using them for a post teaching advanced Italian at a university would be even less valid.

A common framework for validity is Weir's (2005) socio-cognitive validity model, in which different components of the testing process are evaluated: characteristics of the test taker, the test content, the way in which the responses are scored, how well the responses reflect the criterion, and the consequences of the test use. This approach aligns with the examination qualities outlined in the ALTE *Principles of Good Practice* (2020): content validity, construct validity, reliability, criterion-related evidence, fairness, quality of service, practicality, and impact. A discussion of these qualities is beyond the scope of this Manual; please refer to the *Principles of Good Practice* for more information. The key idea to bear in mind in working through this Manual is that every aspect of test development,

administration, rating, scoring, and stakeholder interaction contributes to the validity of a test's score use. An important corollary to this is that validation is not just done once: when revisions are made to item specifications, rating scales, administration instructions, score use, or any other aspect of the test (see Chapter 6), it is important to ask whether the revisions have affected the validity of the test's score use.

1.4.2 Building a validity argument

All these approaches lend themselves to the concept of a validity argument, which lays out claims for the appropriate use of test results and evidence to support that use. As described in the ALTE *Principles of Good Practice* (2020, pp.24–25), the argumentation structure involves making claims about each parameter, providing information to support the claims in the form of explanations, justification based on relevant language testing theory, and backing this up with appropriate evidence. This argument is not necessarily intended to prove that any particular claim is true, but to document that the group responsible for the test has followed best practice and industry best practice standards, supporting the claims about the test score inferences with quality evidence. A validity argument is the central document where the testing organisation stores and updates evidence to support their claim that the test is able to group test takers meaningfully according to their abilities with respect to characteristics that are useful for score users, and is the basis for giving stakeholders confidence in using the test. As noted above, these are living documents that should be updated when changes are made to the specifications, administration, or score use.

1.4.3 A practical example of a validity argument framework: The ALTE Minimum Standards

There are a number of approaches in common use for building validity arguments, many of which may seem daunting because they cover so many aspects of the test (see, for example, Chapelle, 1999; Kane, 2013). The ALTE Minimum Standards (ALTE, 2024) provide a framework for building validity arguments that breaks the elements of such an argument into components specific to language testing so that practitioners have concrete guidelines to follow (see the ILTA Guidelines for Practice for a similar approach). In this Manual, we use the ALTE Minimum Standards (MS) as a way of making the abstractness of the validity argument more concrete in the hope of making it clear what activities testing organisations should pursue. There are 18 MS, each of which provides a high-level reminder of evidence which should be considered, researched and published (where appropriate). These MS cover all stages of the language testing process, though the way they divide the stages differs from how they are depicted in the Testing Cycle used here (Figure 1). Bear in mind that there is no one 'correct' way to distinguish the different kinds of work that go into building a test.

> Note that the term 'standard' is applied in several ways in testing. The MS apply to the quality of the test, much as there are standards for product quality in other fields. 'Standard' is also commonly used to apply to passing criteria, where 'standard setting' refers to the process of establishing a score corresponding to a proficiency standard that test takers must meet (See Sections 2.4 and 5.3). In both cases, 'standard' means criteria that must be met to be considered successful.

The 18 MS are organised into five categories:

1. **Test Construction** (MS 1–5)
2. **Administration & Logistics** (MS 6–10)
3. **Marking & Grading** (MS 11–12)
4. **Test Analysis** (MS 13–14)
5. **Communication with Stakeholders** (MS 15–18)

See Appendix I for examples of practices organisations might undertake in order to provide validity evidence in these areas. This framework is analogous to the CEFR itself: some MS will be more relevant to a given testing organisation's situation than others, but taken together they provide a comprehensive guide to the areas the testing organisation should consider. The key is to determine what evidence is needed for the organisation's specific context and who needs to evaluate the evidence. As stated in the ALTE *Principles of Good Practice* (2020), the aim is to be able 'to make a formal, validated claim that a particular test, or suite of tests, has a quality profile appropriate to the context and use of the test' (p. 24).

1.5 Cross-cutting considerations

In this section, we present general considerations that underlie the stages of the Testing Cycle and the overall validity of score use; understanding these considerations is important for asking useful questions and establishing appropriate activities in support of testing programmes.

1.5.1 Ethics and fairness: The impact of the test on stakeholders and society

Tests can have an impact that concerns a much wider context than the test taker and the institution that carries out the test (see McNamara and Roever (2006b) and the ALTE *Principles of Good Practice* (2020) on the place of language testing in society). Test providers must be aware of this and take it into account when advising other stakeholders. Test impact has several dimensions; in this section we address ethics and validity, with specific attention to fairness, justice, and data protection.

Underlying any discussion of test impact is the concept of the stakes of a test. There is a general understanding that high-stakes tests are those with serious and far-reaching consequences, whereas low-stakes tests have less impact. However, it is important to recognise that there is no clear definition of high or low stakes as a characteristic of a test, and that the use of a test may have different stakes for different stakeholders. See Section 2.3 for a discussion of test purpose and stakeholders.

1.5.1.1 Ethical concerns

Any human activity that affects the lives of others carries with it ethical obligations. The concept of beneficence, the ethical principle of acting to benefit others and prevent harm, is central to considerations of testing impact. Since the early 1980s, ethical concerns have 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. ALTE's LAMI (Language Assessment for Migration and Integration) Special Interest Group (SIG) produced its guideline *Language tests for access, integration and citizenship: An outline for policy makers* on behalf of the Council of Europe and published it in 2016. 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 Chapter 4 and Appendix III).

1.5.1.2 Social consequences of testing as an indicator of possible invalidity

Messick (1989a) draws attention to possible unintended side effects of testing, which will only become visible when attention is directed not only on the test itself but on the societal context in which it is set. These side effects may affect the institution where the testing takes place, such as 'teaching to the test' or 'learning for the test,' or society as a whole when decisions are made based on a test with a questionable construct definition which may not reflect the intention of the test user, or on a test which is related to its construct in a loose manner, thus testing features that were not intended to be tested. Test providers are urged to watch out for negative social consequences of test use as indicators of possible invalidity, and to check whether these are due to factors of test design

and implementation. If negative consequences cannot be traced back to such factors, however, 'the validity of the test use is not overturned' (Messick 1989a, p.88) and possible solutions are to be sought in the political sphere.

Kane, Bachman, Weir, and others took this approach of taking a wider perspective on testing further, with Bachman (2005) systematising the relationship between test validity and test use in his test use argument, and Weir (2005) incorporating test use into his socio-cognitive validity model (see Section 1.4). The key point from all this is that the validity of test score use applies to the intended use as set forth in the specifications and stakeholder communications (see Chapter 2); however, we all recognise that negative consequences come about through uses that were not intended. In such cases, the intended uses are still valid; however, such a situation does still pose an ethical issue that test providers need to address.

1.5.1.3 Fairness

An objective for all test and examination providers is to make their test as fair as possible, meaning that the test content, administration, and scoring do not favour any one group over another for reasons that do not have to do with the construct. (A common example is making sure that the test does not assume familiarity with cultural customs not shared by all test takers.) This is both an ethical obligation and a matter of validity: unfairness violates the principle of beneficence, and it also introduces measurement error (see Section 1.5.2), in that if test takers are advantaged or disadvantaged due to factors unrelated to the test construct, the test is not representing the construct as well as it could. See the ALTE *Principles of Good Practice* (ALTE, 2020), the Code of Fair Testing Practices in Education (JCTP, 2004), and the Standards for Educational and Psychological Testing (AERA et al., 2014). Various other bodies have produced Codes of Practice or Codes of Fairness to assist test providers in the practical aspects of ensuring tests are fair. Kunnan's Test Fairness Framework (Kunnan, 2004) focuses on five aspects of language assessment which need to be addressed to achieve fairness: validity (see Section 1.4), absence of bias (see Chapter 3 for reviewing for bias and Appendix III for statistical measurement of bias), access, administration (see Chapter 4), and social consequences (see Section 1.5.1.2). We discuss bias and access below, as they are not covered in other sections.

Bias means that certain groups of test takers do better (or worse) on the test without this being justified by their different linguistic abilities. This may be caused by some specific factual knowledge which the test requires but which is not shared by all test takers, or by introducing topics that are emotionally charged for some test takers and which thus distract them from performing at their ability.

Access to knowledge about the test and to the test itself is another area in which test performance may differ in ways unrelated to the construct. Differing degrees of familiarity with the test format should not influence the test result; therefore all potential test takers must have the means to inform themselves on test format and requirements for passing the test. The same goes for information on how the test is scored, and how reliable the scores are.

As for access to the test itself, groups of test takers should not be disadvantaged by difficulties travelling to test sites, accessing digital testing platforms, or the like. Fair testing conditions must also be found for test takers with special needs, for example those with limited mobility. Consider also test takers who may be unable to access sections of the test because they have limitations of hearing or sight, or because they are illiterate. Depending on the score use, such test takers may need to be exempted from sections of the test, or there may be accommodations that will enable such test takers to interact with the test as they would with similar materials in real life. Determining a systematic approach to such cases is an essential discussion to have with respect to validity, fairness, and ethics, so that all test takers with similar needs are treated similarly.

1.5.1.4 Justice

McNamara and Ryan (2011) further distinguish between fairness (the impact of the contents and scoring of the test on test takers) and justice (the impact of the existence and use of the test on society) as components of validity. Thus, a test should have a positive social impact, not only by ensuring that the way the construct is measured is free from bias (see above on fairness), but also by ensuring that the construct itself promotes social values of justice, and that the test is used in a way to promote justice. Testing organisations therefore have an obligation to

monitor how stakeholders are using their tests and to take action if they see the tests are being misused.

1.5.1.5 Data protection issues

As testing moves into the digital sphere, test takers produce a large amount of data that may be of interest for various other stakeholders and possibly third parties. As a consequence of this, questions of data minimisation according to the European General Data Protection Regulation (GDPR), data protection, data security, and the right to privacy of personal data become more important as a legal and ethical matter. Consider what data will need to be gathered, processed and stored with relation to the specific test and how the testing organisation can ensure the right to privacy of personal data. Set up a plan for the use of data for (a) validity research for the test, (b) general research conducted by the testing organisation, and (c) research conducted by third parties. Make this plan transparent to test takers, raters and other persons concerned. Obtain consent for the use of any data that will not be anonymised.

1.5.1.6 Ethics, fairness, and justice as part of the validity argument

As is clear in this section, fairness is a key component of validity. As stated in the ALTE *Principles of Good Practice* (2020),'fairness is the third fundamental consideration accompanying validity and reliability' (p. 20). Key ethics, fairness, and justice principles that need to be addressed in any validity argument are:

- All test takers have equal opportunity to demonstrate ability
- Scores mean the same thing regardless of test takers' background
- Clear information is accessible to all
- Appeals and accommodations processes are transparent
- Regular monitoring detects and corrects bias
- The best interest of test takers is a guiding principle

1.5.2 Ensuring reliability throughout

Reliability in testing means consistency: a test with reliable scores produces the same or similar result on repeated use for a test taker of the same ability. This means that a test would rank-order test takers in nearly the same way and would lead to similar decisions concerning test takers from one administration to another. As stated in the ALTE *Principles of Good Practice* (2020), 'reliability concerns the extent to which test results and feedback are precise, stable, consistent, and free from errors of measurement' (p. 18). Reliability is a prerequisite of validity, because we cannot put any trust in unstable or unreliable test results. Note, however, 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.

Test scores vary between test takers. This is because test takers have different abilities, but some variation may also be due to factors unrelated to their ability, such as features of the test room (noise, warmth, others), the test taker's health/fitness, differing performance of raters or differing difficulty of the test version. Reliability is defined as the proportion of test score variability which is **caused by the ability measured**. 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. The test developer should be aware of the likely sources of error, and do what is possible to minimise it. Following the procedures in this Manual will help. The ALTE *Principles of Good Practice* and Minimum Standards also address reliability at multiple points.

Following good practices can improve reliability, but how can we measure the reliability of a test? See Appendix III for common statistical methods. Note that there can be no universal 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 people who have already passed some selection procedure and who are, as a consequence, closer to each other in ability, 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 method of marking. The scores of rated tasks, such as a writing task, are typically less reliable than those for objectively marked items, such as multiple-choice items, because more variance (error) is introduced in the rating process than in an objective scoring process. This, however, does not mean that objectively marked items are to be preferred over rated tasks. For rated tasks, reliability can be much enhanced by rater training, or the use of multiple raters. See Chapter 5, Section 5.2, for a discussion of rating.

Given these sources of variation in reliability, how can a target reliability be defined?

Studying reliability on a routine basis will be useful in identifying particular test versions which have worked better or worse, and in monitoring the improved quality of tests over time. Routine reports will reveal the reliability that can be expected for a particular test given its length, test taker population, raters and item/task types. The higher the stakes of the test, the higher the reliability needs to be. See Appendix III for discussions of statistical reliability.

Most reliability estimates, such as Cronbach's Alpha, the Pearson correlation, 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 as acceptable, and above 0.8 as desirable. However, measures of reliability that compare the ratings of two raters can actually be too high: values of 0.95 might indicate that the raters are not rating independently, which can pose a different kind of problem.

For smaller testing programmes where statistical reliability estimates may not be feasible, ensure that good practices are followed for test development, administration, and rating, and consider methods for mitigating the effects of unreliability on test score use:

- Rely on well-founded rating systems with extensive standardisation
- Focus on rater training and monitoring
- Consider what existing evidence in addition to the test can contribute to the decisions to be made
- Develop portfolios or continuous assessment practices to provide additional information

1.5.3 Practical constraints and solutions

It may happen that circumstances will not permit the implementation of some of the procedures outlined in this Manual. If that is the case, it is important to think of alternative ways to ensure the test's validity, reliability and fairness. Consider what actions are possible that will show that the test scores provide useful information and are properly interpreted. Common challenges are a small number of test takers and limited resources as to financing and staff. Below are a few suggestions on how test quality can be maintained under those common circumstances. Please note that they are only examples to demonstrate possible alternatives, and they are not appropriate for all contexts. Each test provider must determine procedures appropriate for their context.

Small test taker numbers will typically mean that some statistical validation measures cannot be carried out because the results will lack significance (i.e. they may be due to chance; see Appendix III on sample size). When designing a new test which is not expected to be taken by many test takers, take this into account right from the beginning. Some approaches that may be helpful are discussed below. They are:

- Item or task types that lend themselves to qualitative evaluation
- Gathering qualitative evidence during piloting or trialling
- Piloting and trialling items and tasks under testing conditions
- Evaluation of rating criteria

When identifying item and task types to be used (see Chapter 2), focus on **item/task types that lend themselves to qualitative evaluation;** that is, instead of multiple-choice or multiple-matching items which typically require large test taker numbers for validation, focus on productive tasks like interviews, answering to prompts, short answer items or essay tasks. For a test that already exists, it also makes sense to **focus on qualitative rather than quantitative evidence**. For instance, in order to get at the difficulty and clarity of items and tasks, the testing organisation may conduct

(standardised) interviews with a limited number of piloting or trialling test takers or examiners, or have piloting test takers think aloud while solving the tasks, instructing them to focus on features that are particularly difficult or easy, or ambiguous, and record this (See Section 3.3.3). In order to ensure that items and tasks are clear and unambiguous, it may also be desirable to **pilot or trial the test** (see Chapter 3) with a limited number of accomplished speakers of the language of the test, and look for discrepancies. **Ensure that the rating criteria are clear**. Rating idiosyncrasies or **diverging interpretations of the criteria** can be identified by having raters rate the same samples, and discussing the results, followed by any needed revisions of the rating criteria or training materials. Please note that these methods are not necessarily a second-best choice. Both quantitative and qualitative methods are important for establishing validity. It is the balance and emphasis on qualitative and quantitative methods that may differ depending on population size: large testing programmes may rely more on quantitative data because for them it is more reliable and easier to gather, but they will typically include some qualitative methods. Where quantitative work is not reliable enough, whether because of small numbers of test takers or because the reliability analyses revealed problems, there needs to be more qualitative work.

Even with small programmes, quantitative analysis is sometimes possible: it may make sense to collect data over multiple administrations in order to reach enough test takers to run quantitative analyses (see Appendix III for the relationship of test taker numbers and significance).

Small numbers may also make it hard to collect enough performance samples to conduct a standard setting with (i.e. to establish a cut-off score on a test component or on the whole test, see Chapter 5.3), or to find enough experts to judge the performance samples. Please consult *Aligning Language Education with the CEFR* (British Council et al., 2022) and Section 5.3 for ways to cope with this situation.

In order to save effort, it may be practical to use existing tests or procedures as a model or at least as a starting point and in this way benefit from other institutions' experience. Regardless of any practical limitations, there may however still be questions about the test's validity that it will not be possible to answer. It is good practice to be transparent about such limitations. Stakeholders, especially the groups requesting the test, may be able to help when they are made aware of such problems. In any case, be sure to implement a continuous improvement procedure with regular reviews of the test (see Section 1.6.3 and Chapter 6).

1.6 Planning and managing test development

This section pulls together the cross-cutting concepts discussed so far and outlines how they are applied across the Testing Cycle (Figure 1) to activities needed for a validity argument (Section 1.4.2). For more detail on the activities, see the relevant chapters in this Manual and the commentary on the Minimum Standards in Appendix I.

1.6.1 Creating a comprehensive validation plan

Every test requires a validation plan, but the scope and depth of that plan will vary according to the stakes and use of the test. An in-class progress test designed by a single teacher, for example, may need to focus on alignment with the curriculum and how well the test predicts performance in the next term, with limited evidence about staffing and training, while an international high-stakes admissions test must generate comprehensive documentation across all stages of the Testing Cycle (see Figure 1). A good plan therefore acts as the blueprint for validation: it identifies what claims will be made, what evidence will be gathered, when and how this evidence will be reviewed, and how it links to the evaluation standards to be used. In practical terms, the validation plan is the first layer of quality assurance for a new test: it organises resources, sets milestones, and ensures that subsequent documentation (see 1.6.2) and continuous improvement activities (see 1.6.3) are embedded from the outset. The elements of validation planning for a new test are shown below. Validation plans can also be applied to existing tests, as a part of continuous improvement. For these tests, the steps need not be done in the order shown below; testing programmes will typically begin with the activities listed under Operational Test Use, and then determine priorities for applying

validation activities to the other elements. Note also that these elements of validation planning, while generally aligned with the Testing Cycle, are designed to produce evidence that information from each stage of the cycle is brought to bear on other stages so that the test as a whole is functioning properly. The elements below contain links to the ALTE Minimum Standards (MS), for ease of reference to Appendix I, but they can be used with any evaluation standards.

1. **Initial planning (Pre-development)**

- Define purpose and population (MS 1) – good practice often involves clarifying the decisions to be supported, the TLU domains, users, stakes, and any exclusion criteria (see Chapter 2 on defining the construct and specifications).
- Establish theoretical model and framework (MS 2) – many programmes find it helpful to identify the model(s) of language use (e.g. CEFR) and show how they will be operationalised in specifications and scoring (see Chapter 2 on the construct and frameworks).
- Design quality management system (MS 3) – this can include setting ownership, approvals, risk/change control, and document controls (versioning, retention). General management principles run through all chapters; see especially Chapter 3 on quality assurance and Chapter 6 on communication.
- Allocate resources and timeline (MS 3, MS 6) – good practice is to assign roles, skills, and budget, and a milestone plan covering development, administration, marking, data collection, communication, and operational use (see Chapters 2–3 for timelines on development/piloting and Chapter 6 for stakeholder scheduling).

2. **Development**

- Create and try out test specifications (MS 4) – good practice includes defining construct coverage, blueprints, item/task types, difficulty targets, marking approach, and version equivalence (see Chapter 2 for construct definition and specifications).
- Recruit and train team (MS 3) – programmes often define role profiles, minimum qualifications, and an initial/ongoing training and calibration plan (see Chapter 3 on item writing, reviewing, and test construction). Prepare and train automarkers, if they are used, and define their role and functionality (e.g. clerical marking, rating).
- Develop initial item pool (MS 4–5) – this may involve producing, reviewing, and pre-screening items/tasks against specifications, bias/fairness checks, and content balance (see Chapter 3 on item writing and pre-testing).
- Plan and conduct alignment studies (MS 5) – typical approaches include planning the methods (e.g. standard setting, performance profiles) and defining the evidence required for claims of alignment to particular levels (see Chapter 2 on performance profiles and Chapter 5 on standard setting).

3. **Establishing administration, marking, data collection, and communication procedures**

- Rehearse administration procedures (MS 6–10) – many programmes rehearse delivery, security, accessibility/accommodations, and data protection; logging deviations and fixes are also common (see Chapter 4 on administration).
- Train initial markers, raters, or automarkers (MS 11–12) – good practice is to standardise raters, collect double-marked/double-rated samples, and set monitoring metrics for both human and automated marking and rating (e.g. drift, agreement) (see Chapter 5 on rating and marking).
- Collect preliminary data (MS 13–14) – examples include piloting, pretesting, and/or trialling of items, tasks, and versions, and running item analyses, reliability estimates, DIF/bias screens, and provisional equating where feasible (see Chapter 3 on the use of preliminary data and Chapter 5 on analysis).
- Develop communication materials (MS 15–18) – this can include drafting test taker handbooks, reporting templates, appeals/feedback routes, and public summaries (see Chapters 2 and 6 on communication and stakeholder engagement).

4. **Operational test use**

- Continuous monitoring and improvement – many programmes review evidence routinely to ensure compliance with all MS (see Chapter 5 for analysis and Chapter 6 for monitoring impact/fairness).
- Regular stakeholder engagement – examples include formal consultations, feedback loops, and scheduled updates (see Chapter 6).
- Periodic reviews (internal and external) – good practice is to schedule structured internal reviews at regular intervals (depending on the administration cycle and method and the size of the testing population, this could be after each administration, or once or twice a year) and occasional independent audits for transparency and external assurance (linked to MS 18, and addressed in Chapter 6 on accountability and quality assurance).

A comprehensive validation plan is therefore more than a checklist of activities: it is the structured roadmap that connects the construct, the evaluation standards, and the Testing Cycle (Figure 1). By mapping evidence needs to specific stages, and by cross-referencing them with the detailed guidance in later chapters, programmes can demonstrate that validity claims are well founded and transparent to stakeholders. The plan also provides the framework within which subsequent documentation and evidence (see 1.6.2) can be collected, organised, and reviewed, making it the foundation for a defensible and sustainable testing programme.

1.6.2 Documentation and evidence

Documentation is the thread that connects the entire Testing Cycle: it transforms intentions into traceable actions, and actions into evidence that supports validity claims. A well-maintained evidence base ensures that decisions taken during test design (see 1.6.1), alignment with the construct (see 1.3.3), and the requirements of the evaluation standards are visible, auditable, and defensible. Systematic documentation also enables continuity: teams change, but a transparent record of procedures, rationales, and evidence allows future colleagues, both internal and external, to understand why particular choices were made and how the test has evolved. In short, documentation is not an administrative afterthought, but an essential quality-management practice that underpins validation, facilitates stakeholder trust, and supports continuous improvement (see 1.6.3).

1.6.2.1 Types of documentation

Typical useful documents include:

- **Policy documents stating intentions** – e.g. purpose, fairness/accessibility, data protection, security, and appropriate-use statements. See Section 1.5 on fairness, ethics, and data protection, Chapter 2 on purpose and construct, and Chapters 2 and 6 on stakeholder communication.
- **Specifications detailing the requirements for test construction** – typically the specifications document that is the subject of Chapter 2.
- **Procedural documents showing implementation** – e.g. item writing and review manuals, administration handbooks, rater guides, and grading or appeals procedures. See Chapter 3 on item writing, Chapter 4 on administration, and Chapter 5 on marking.
- **Evidence documents demonstrating effectiveness** – e.g. trial/pretest and live datasets, item statistics, reliability/equating/DIF reports, and standard-setting outputs. See Chapter 5 on analysis, reliability, and alignment.
- **Review documents for planning improvements** – e.g. risk registers, incident logs, change request and decision records, and annual technical reports. See Section 1.6.3 on continuous improvement and Chapter 6 on accountability and quality assurance.

1.6.2.2 Documentation management

To keep track of documentation and ensure that the appropriate documents can be readily retrieved, some system of documentation management is useful. Large testing programmes will typically have

dedicated systems for this, but even for smaller programmes, good practices can be followed with standard office processes and software. Documentation management tools include:

- **Version control system** – e.g. unique IDs, semantic versioning, approvals, and release notes. Relevant across all chapters; see especially Section 1.6.1 on planning and Chapters 3 and 6 on quality management.
- **Regular review cycles** – e.g. scheduled checks (per administration, quarterly, annually) with clear owners and deadlines. See Section 1.6.3 on continuous improvement.
- **Accessible storage and retrieval** – e.g. role-based access, indexed repositories, and consistent file naming with metadata. See Chapter 3 on managing materials, Chapter 4 on data protection, and Chapter 6 on communication.
- **Clear responsibility assignments** – e.g. RACI (Responsible, Accountable, Consulted and Informed) models for each document/evidence stream and defined handover points. See Chapter 3 for team roles.
- **Audit trails for changes** – e.g. record who changed what, when, and why, and link each change to supporting evidence and communications. See Chapter 3 on managing materials and Chapter 6 on accountability and transparency.

Ultimately, documentation and evidence provide the backbone of a defensible validation argument: they ensure that claims made about test use are transparent, replicable, and grounded in verifiable practice. Without a structured record, even well-designed procedures risk being forgotten, misapplied, or questioned by stakeholders. With it, every stage of the Testing Cycle can be traced, reviewed, and, where necessary, improved. In this way, documentation not only fulfills compliance with the evaluation standards but also creates the foundation for the continuous improvement cycle described in Section 1.6.3.

1.6.3 Continuous improvement cycle

Continuous improvement is the mechanism that keeps a test programme both valid and responsive over time. Even when a test has been carefully planned (see 1.6.1) and thoroughly documented (see 1.6.2), evidence from live administrations will reveal new challenges, risks, and opportunities. No test can remain static: populations shift, delivery modes evolve, and stakeholder expectations change. As illustrated in Figure 1, the Testing Cycle, each stage generates feedback that should loop back into earlier stages, sometimes even requiring a revision of the construct or specifications, which in turn inform all stages of the cycle. Embedding a structured improvement cycle ensures that this feedback, whether psychometric evidence, rater monitoring, stakeholder feedback, or internal/external reviews, is systematically analysed, acted upon, and integrated into the validation argument documentation. In this way, continuous improvement both supports compliance with the evaluation standards and demonstrates a visible commitment to reliability, fairness, justice, ethics, and social responsibility. A common framework for improvement is the Plan-Do-Check-Act framework:

1. **Plan**: Review MS requirements and current status – set targets/thresholds, prioritise risks, and define the research agenda. See Chapter 2 for specifications and constructs; Chapter 6 for risk management and governance.
2. **Do**: Implement improvements – try out changes when possible, update training/materials, and communicate changes to stakeholders. See Chapter 3 for item/test development processes; Chapter 4 for delivery and training; Chapter 6 for stakeholder communication.
3. **Check**: Gather evidence of effectiveness – monitor key performance indicators (e.g. reliability, rater drift, item fit/discrimination, DIF, complaints) and impact/fairness. See Chapter 5 for statistical analyses and bias checks; Chapter 6 for fairness and impact monitoring.
4. **Act**: Institutionalise successful changes – adopt or roll back with documented rationale; update specifications, policies, and the validity argument. See Chapter 2 for updating specifications; Chapter 3 for revisions to content; Chapter 6 for updating policies, reporting, and accountability.

Although many opportunities for improvement may arise organically, as responses to problems detected through the normal working processes or to changes in the testing population or environment, it is important to set aside time for a comprehensive review at regular intervals to ensure that all aspects of the testing programme receive attention. This review often takes place annually, but depending on the particular situation, it may make sense to review at the end of each academic term, or after each major administration, or after an item pool is refreshed. Elements of the regular review include:

- Stakeholder feedback analysis – synthesise test taker/institution feedback and appeals data into actionable themes. See Chapter 6 on stakeholder engagement and appeals handling.
- Performance indicator review – compare results against targets; investigate outliers and recurring issues. See Chapter 5 on psychometric monitoring and reporting.
- Benchmark against best practices – scan sector standards/guidance and similar programmes for improvements. See Chapter 6 on accountability and external review.
- Priority setting for improvements – rank by risk, impact, feasibility; assign owners and timelines. See Chapter 6 on governance and quality assurance frameworks.
- Resource allocation decisions – align budget and staffing with the improvement plan and scheduled studies. See Chapter 3 on team competence and resourcing; Chapter 6 on institutional planning.

Taken together, planning, documentation, and continuous improvement form a cycle that underpins the quality of any testing programme. The next section (1.7) invites each testing organisation to reflect on how well these principles are embedded in their own context, and to identify areas where further evidence or action may be required.

1.7 Key questions

Having reviewed the principles of validation, documentation, and continuous improvement, it is useful to pause and consider how these apply to the particular testing programme of interest. The following questions are designed as prompts: they do not prescribe a single approach, but rather help identify where further clarification, evidence, or development may be needed. Working through them will also prepare organisations for the more detailed guidance in the chapters that follow, where each stage of the Testing Cycle (see Figure 1) is explored in depth. These questions are meant for the testing organisation as a whole, and may need to be answered by different people within the organisation.

- **What is your test purpose?** (MS 1; see Chapter 2 on the decision to provide a test and planning)
- **How does your test purpose align with stakeholder needs?** (MS 1; see Chapter 2 on defining the construct and specifications)
- **What is your model of language competence?** (MS 2; see Chapter 2 on theoretical frameworks and the CEFR)
- **Is the language model you use clearly operationalised in your specifications and processes?** (MS 2; see Chapter 1 on models of language use and Chapter 2 on operationalising frameworks)
- **Are your team qualifications and training documented?** (MS 3; see Chapter 3 on team competence and item writing)
- **How do you ensure version equivalence?** (MS 4; see Chapter 2 on specifications and Chapter 3 on constructing test versions)
- **What evidence supports your claims about score interpretation?** (MS 5; see Chapter 5 on alignment and standard setting)
- **Are your administration procedures standardised?** (MS 6–8; see Chapter 4 on administration and delivery)
- **How do you protect test taker data?** (MS 9; see Chapter 4 on administration, security, and data management)

- **What accommodations do you provide?** (MS 10; see Chapter 4 on accessibility and fairness)
- **How reliable is your marking?** (MS 11–12; see Chapter 4 on marking procedures and Chapter 5 on monitoring rater reliability)
- **What kind of bias is most relevant to your situation? How do you check for it?** (MS 13; see Chapter 5 on analysis and fairness checks)
- **Do your items function as intended?** (MS 14; see Chapter 5 on item analysis)
- **How quickly do you report results?** (MS 15; see Chapter 6 on reporting)
- **What information do stakeholders need?** (MS 16–17; see Chapter 6 on communication and transparency)
- **How do you maintain stakeholder engagement and ensure that your test scores are appropriately used?** (MS 18; see Chapter 6 on stakeholder management and social responsibility)

By reflecting on these questions, those responsible for the test will build a clearer picture of the current strengths and potential gaps in the testing programme. This reflective baseline provides a foundation for the practical steps discussed in the next chapter, where we approach the Testing Cycle with the central task of defining the construct and writing effective test specifications.

1.8 Further reading

Practical manuals

There are many works with guidelines and advice for language testing practitioners that are referenced throughout this Manual. For general guides to good practice, see ALTE (2020), EALTA (2006), ILTA (2007), and Bachman and Palmer (1996). For CEFR alignment, see Council of Europe (CEFR 2001, 2009, CEFR Companion volume) and ALTE et al. (2022).

Foundational background topics

Language proficiency

CEFR 2001 and the CEFR Companion volume provide extensive discussion of the model of language proficiency. Of particular interest is the discussion of profiles (CEFR Companion volume, pp.38–40), articulating the degree of independence of the descriptors within and across categories. The ALTE Language Assessment for Migration and Integration Special Interest Group (LAMI SIG) published a report on 'uneven profiles' (ALTE, 2023) that complements the discussion in the CEFR Companion volume.

Validity

Messick (1989a), Weir (2005), and Kane (2010) are fundamental works that all language testers should be familiar with. On the concept of validity arguments, see Bachman (2005) and Chapelle and Voss (2021). ALTE's *Principles of Good Practice* (2020) also provide a good summary of validity concepts.

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.

Impact

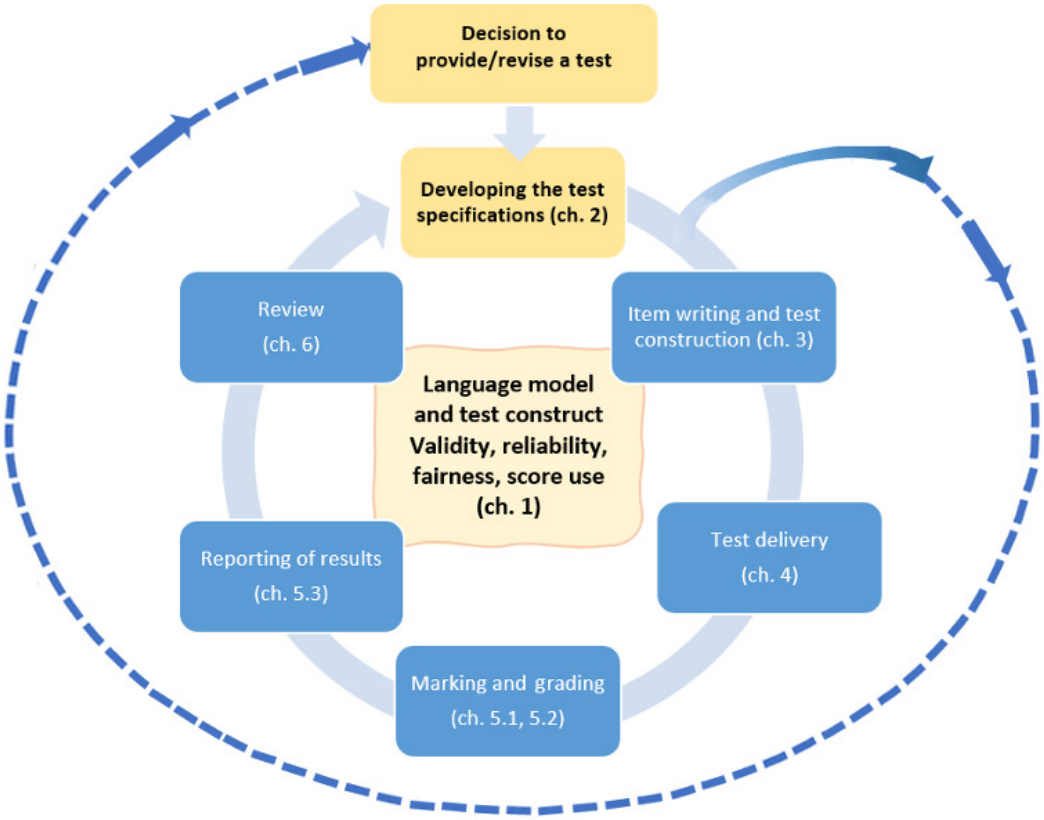
McNamara and Roever (2006b) is an important resource for the place of testing in society, and it includes discussions of fairness and ethics. Kunnan (2004) also provides good conceptual foundations for approaches to fairness. For impact issues specific to testing for immigration and citizenship, see ALTE (2016). For ethics, there are several useful codes of ethics for language testing, notably ILTA (2000) and EALTA (2006).

Next steps

Chapter 1 set the stage by articulating the stages of the Testing Cycle and introducing fundamental underpinnings of language test quality: the concept of a construct, including discussion of language models and the CEFR framework; how to build a validity argument; essential considerations of score use, impact, fairness, ethics, and justice; and practical guidance on testing processes and documentation.

Chapter 2 will apply these fundamental considerations to the development of test specifications, starting with the decision to develop a test and the planning phase. The central processes described in the chapter will be related to the development of construct and specifications. The relation with the stakeholders will be analysed and presented in connection with every phase in the stage.

2 Developing the test specifications



2.1 The process of developing the test specifications

The aim of this stage is to define the construct and how it will be operationalised through test specifications which can then be used to construct tests. The information they include mainly refers to

- The test content – what is tested
- The test format – e.g. test components, type and number of tasks and items
- Marking and rating guidelines and instruments

The specification development stage begins when the test provider or the test sponsor decides that a new test is necessary or that a test needs to be revised.

Figure 3 illustrates the process of developing specifications, showing the three essential phases (planning, design and try-out) underpinned by the process of cooperation with stakeholders, which is relevant to all the steps of specification development.

1) The final specifications are the output of the development stage. However, the path to this outcome may not always be straightforward. Findings from the design or try-out phase (e.g. resulting from consultation with stakeholders) may necessitate a return to the planning phase. From try-out the test developers may be taken back to design. In any of these cases, some steps would need to be repeated.

2) Even after the specifications are finalised and put into use in developing a test, information from various stages of the Testing Cycle may cause the specifications to be revised. This situation might occur especially with tests newly put in use. See Chapter 6 for information on test revision.

2.2 The decision to provide or to revise a test

Although the decision to provide a test is not considered part of the specification development process outlined in this Manual, it provides important input into the planning phase, as the requirements and background information identified by the test provider or the test sponsor will largely determine the design of the test and how it is used. A significant revision to an existing test may also be decided, for example, in the case of transitioning from paper-based to digital delivery or of incorporating automated marking. Such a decision may impact all or most phases in test development. See Chapter 6 for information on the revision of the test.

The decision to develop a new language test or to revise an existing one may be driven by different stakeholders. Sometimes, the test provider makes this decision and carries out the test development project. In other cases, a third-party sponsor commissions the test provider to develop or revise the test. A new (or revised) test can have significant implications for test takers and other stakeholders, for example in contexts such as migration or admission to higher education. Test providers have great social and ethical responsibilities regarding the test results and how these are used. They must make the test sponsor aware of the suitability and possible consequences of certain decisions related to the test (e.g. regarding the CEFR level required for the labour market or for a pre-entry test). The necessity of the test itself may be called into question. The test provider may have reservations about the appropriateness of the test and suggest alternatives. For example, language and Knowledge of Society (KoS) courses may be more suitable for migrants and refugees than a corresponding test, in certain contexts. Discussions between the test provider and the test sponsor must address the issue of how to avoid test misuse at all stages of test development and administration. In the specifications development stage, misuse needs to be prevented; once the test becomes operational, misuse needs to be detected and removed or mitigated. Communication between the test developer and test sponsor should be ongoing throughout the stages, as issues may arise at a later stage that raise questions about a previous stage.

The test requirements and background information must be clearly identifiable. The next section provides information on how to do this.

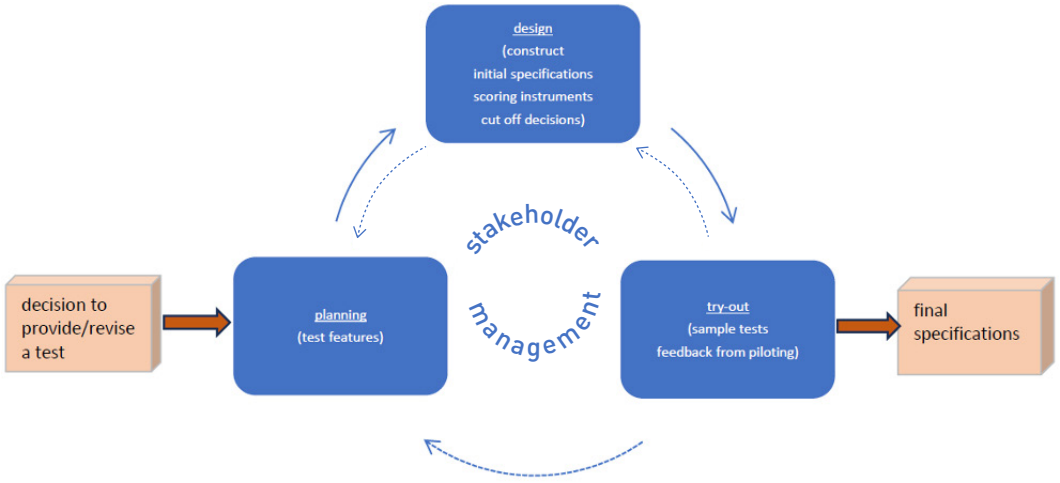


Figure 3 ***The process of specification development***

2.3 Planning

This phase is dedicated to gathering information for use in later stages (e.g. about the characteristics of the test takers, the purpose of the test, the possible relationship to a context – educational, migrational, occupational, etc.). Much of this information should come from either the test provider or the test sponsor – whichever party decides that the new test/the test revision is necessary. However, a wider range of stakeholders should be consulted, including, for example, ministries and government bodies, publishers, schools, parents, experts, employers, educational institutions and administrative centres, migration regulators and migration support centres, and human rights organisations. Where many informants are involved, structured information may be collected using questionnaires, interviews, focus groups or seminars. In a classroom situation, however, direct personal knowledge of the context and the test takers may be sufficient. For all tests, but particularly those related to language for specific purposes (LSP), it may be necessary to gather more precise and focused information on the language needs of learners/test takers. Consequently, a needs analysis (NA) process is strongly recommended. As part of this process, information on the domain and tasks will be gathered through clear procedures and with the help of stakeholders. The characteristics identified in the target language use (TLU) through the needs analysis process should then help the test developer design tasks with similar characteristics.

The test developer needs to clarify a series of questions in the planning phase. The information they obtain will help design the specifications and implement the test versions. The most important of these questions are:

Questions on the test taking population, test purpose and context

- What are the characteristics of the test takers to be tested? (age, gender, social situation, educational situation, L1, plurilingual profile, pluricultural profile, literacy profile, digital skills, access to technology, etc.). How are special needs accommodated? (e.g. visual, hearing, mobility limitations, motor control limitations, learning disabilities)
- What is the purpose of the test? (school-leaving certificate, admission to a programme of education, minimum professional requirements, formative or diagnostic function, migration or citizen-related purposes, etc.)
- Does the test relate to an educational context (a national or regional language policy, a curriculum, a methodological approach, learning objectives, etc.), a migration or citizenship context (e.g. emphasis on oral vs. written communication, sensitivity concerns around refugees), a professional/occupational context, etc.?
- What standard is needed for the proposed purpose? (in terms of CEFR levels and competences, a standard relating to a specific domain, etc.)? How many levels of the standard will be assessed?
- If multiple levels need to be tested, will it be more effective to test them as a series of single-level tests, as a single multi-level test, or as an adaptive test?
- What sources of bias might there be? (e.g. in a test designed for a homogeneous population, references to cultural practices of the population would not present a problem, whereas they might be a source of bias if the test is given to a more diverse population). How can we make sure that bias and fairness are addressed in test content, the development process and test delivery?

Questions on the use and impact of results

- How will the test results be used? How will you make sure to capture all of the appropriate uses of the test?
- Who are the stakeholders and how will appropriate uses of test results be communicated to them?
- How can misuse of test results be prevented? (e.g. design of score reports to include caveats and appropriate interpretations, tasks on the test that are obviously related to appropriate use). How will misuse of test results be detected and addressed once the test is operational?

- What kind of impact can be expected (see Section 1.5.1) and how will it be monitored? (e.g. at an institutional, social level)
- How will the test performance be monitored over the long term? (test data analysis, feedback collection from the stakeholders involved in the test development and administration, test taker input, score user input)

Questions on practical considerations

- How will the test be financed and what is the budget?
- Who will be responsible for each stage of test provision and how will the people be trained for their different roles? (i.e. materials production and test construction, administration, proctoring, setting cut-off scores, marking, grading, reporting of results, monitoring and revision)?
- How will pretesting, piloting or trialling be conducted (see Section 3.3.3)?
- When should the test be ready?
- How many test takers are expected? How many members of staff and with what qualifications will be needed?
- Will the test be administered in scheduled sessions? If yes, how often? Or will it be available on demand?
- What mode of delivery is required (e.g. paper-based, networked computer, mobile device)? Does the mode of delivery meet the needs of the testing population and other stakeholders? If the test is delivered in multiple modes, for example, on paper and digital, or you are transitioning from one mode of delivery to another, are the tests comparable in content, item/task type, item difficulty, the scores test takers achieve, bias, construct-irrelevant variance, etc.?
- Will it be possible to take only some components of the exam, or is the exam administered only in all its components? Will it be possible for test takers to repeat the test or a component of the test if they are not satisfied with the results?
- Where will the exam be administered (i.e. classroom, school computer lab, testing centre, at home) and what arrangements are possible for accommodating test takers with special needs?
- Will the test be administered only in the country in which it was created, or/also in other countries? Will the test be administered by the test developer and/or by (other) examination centres or test delivery providers?
- Which aspects of the test involve automation (e.g. item writing, delivery, marking, etc.) and what processes are in place to ensure the work done by automated systems is appropriate?
- What level of security is required before, during and after administration and how will it be ensured?
- What data about test takers do you need to collect in order to administer the test? What legal requirements for data collection, sharing, and maintenance need to be met?
- What requirements do stakeholders have about test takers, data collection and retention?
- Do any stakeholders have specific requirements or prohibitions regarding the use of AI for test development, administration, marking, rating and reporting of results?
- Do any stakeholders have specific requirements or prohibitions regarding the possible multiple roles of the persons involved in testing? (e.g. stakeholders might require that test developers cannot also be the markers or persons who taught the test takers cannot administer the test)
- At the end of the planning phase the test developer should have clear information on the test requirements and background information (see the example in Table 1 below for a hypothetical test to be used for migration context). Special attention needs to be given to adapting the features of the test to the group of test takers (e.g. if the test takers are children, if they are low literate).

Table 1 ***Outcome of the planning phase (example)***

purpose and context	test of proficiency in the migrational domain
target population	adult, medium and high level of education, various L1
content	general
targeted skills	oral and written reception, production, and interaction
targeted level	CEFR B1 across the whole test
intended use of results	obtaining citizenship
mode of administration	digital, in local testing centres

2.4 Design

The information collected in the planning phase provides the basis for design. During this phase, the construct is defined, the draft specifications and some sample test materials are created, and the first decisions regarding cut-off scores, marking and rating are made. The first step is to define the construct, which will guide both the design of the test and the interpretation of results.

2.4.1 Defining the construct

Unlike physical attributes such as height, language competences cannot be measured directly. Consequently, we need to define the knowledge and abilities that we want to measure in the test construct, and then operationalise it through specifications, tasks and items. The construct normally relates to the whole Testing Cycle and the process of validation: 1) it relates to a model of language use (e.g. the action-oriented approach of the CEFR); 2) it is operationalised in specifications; 3) it is measured through tasks and items and marking procedures which are built based on specifications; 4) the competences measured through the construct are referred to when the results are reported; 5) the validation of score use evaluates whether the construct adequately represents the TLU domain.

> If the test is significantly revised, the construct may need to be adapted to meet new needs and objectives. Changes in the construct may also be determined by evolutions in different stages of the Testing Cycle (e.g. transitioning from paper-based to digital delivery or feedback from score users once the test results were reported and are in use). Since the construct is at the centre of the Testing Cycle (see 1.3.3), changes to the construct will require re-evaluation of most aspects of the Testing Cycle. See Chapter 6 for information on test revision.

An important aspect that needs to be included in the definition of the construct is the realm of TLU. This can be defined in general terms (personal, public, occupational and educational, according to CEFR 2001, 4.1.1, Table 5) or more specifically, depending on the purpose of the test (e.g. academic, migrational, for specific purposes – medical, legal, business, etc.). Defining the TLU domain provides a framework for developing the test specifications and, based on them, the tasks.

Once the TLU is defined, the next step is to consider the communicative activities that are important for the test purpose. Will stakeholders need information about test takers' ability to take notes on a lecture, write position papers, compose business emails, or write instructions? Knowing which activities are needed will guide selection of the particular CEFR scales that are needed for the test.

As relevant activities are defined, it is also important to consider the level requirements for each activity. As discussed in the CEFR Companion volume (Section 2.7), different TLUs may require differing abilities for different activities. If a construct identifies different levels for different activities, the testing organisation will need to determine how best to report scores: for many tests, it is unrealistic to expect reliable scores for individual descriptor scales, so a test component will generally combine items and tasks targeting multiple descriptor scales and, potentially, levels.

In defining the construct, the naturally uneven profiles of test takers need to be considered. As indicated in the CEFR Companion volume (Section 2.7), the development of differentiated profiles is encouraged. Identification of relevant activities for a particular group of learners/test takers needs

to be followed by establishing the level the learners/test takers need to achieve in those activities in order to accomplish their goal. This will impact on the content of the test, but also on the design of the rating instruments and on the way results will be reported (See Sections 1.3.2.4, 2.4.4 and 5.4). Consultation with stakeholders is needed already in the planning and design phases in order to clarify the benefits of this approach. Multi-level tests, using a wider range of tasks across levels, can help with profiling the test takers. For example, in a test targeting reading and oral comprehension, displaying items for Levels B1 and B2, the outcome might be reported as a profile, with a test taker being awarded Level B1 in oral comprehension and B2 in reading comprehension. The outcome does not necessarily need to be one of the two levels across the whole test.

The construct will be further developed into the test specifications.

An example could be the construct of a test for the purpose of applying for citizenship. The construct of such a test might include scales for oral and written reception, oral and written production, and oral and written interaction. Figure 4 presents a fragment of a possible construct for the written interaction portion of the construct at Level B1 in this context, based on the CEFR model of language use. This construct is further developed into the corresponding specifications in the following section (2.4.6).

1) This is just an example of a partial possible construct, reflecting the purpose of a particular test. Besides the test requirements and context, it is informed by those illustrative descriptor scales in the CEFR Companion volume which were considered relevant for the context of applying for citizenship. When building the construct of a test or adopting a model of language use as a construct, each test developer must relate to their own context, purpose and target population.
2) Figure 4 represents a scheme of the possible construct, which can be supported by a descriptive, explanatory document detailing the rationale of the construct.

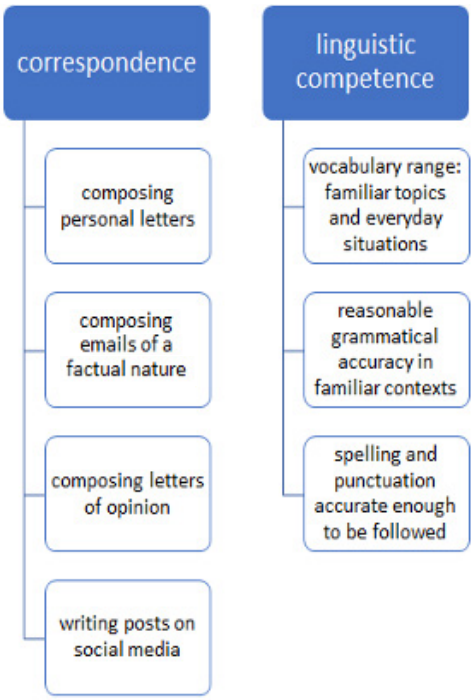


Figure 4 ***Example construct. Partial (written interaction)***

2.4.2 Considerations on test content

A clearer idea of the test content and format needs to be developed in the design phase. 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(s) (e.g. Table 1).

Based on this data, on the model of language use chosen for the test and on the construct, the draft specifications will be built. The test developer will have to consider essential aspects related to test content, such as: the context of language use; the types of tasks which would mirror the ones in TLU; the types of texts used as input or required as response to tasks, sources of texts and prompts.

The CEFR 2001 is a useful resource for defining the characteristics of the test, with many of its chapters relevant to testing. 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 9 on assessment discusses how the CEFR 2001 can be used for different testing purposes. As for information related to the construct, Chapters 4 and 5 of the CEFR 2001 can be consulted along with the enhanced version of Chapter 2 of the CEFR Companion volume, where the important extensions and additions of the Companion volume to the CEFR 2001 groundwork are explained. More focused reference to sections of CEFR will be made below, in relation to test content, along with reference to other relevant sources.

2.4.2.1 The context of language use (domains, situations, conditions and constraints)

As a result of the information gathered in the planning phase and a needs analysis (if one is used), the test will be mainly related to one of the CEFR's domains which define human life: personal, public, occupational, educational.

Chapter 4 in CEFR 2001 refers to the context of language use. More precisely, some indication of topic areas considered suitable for use can be found in Sections 4.1 and 4.2.

For types of real-life situations relevant to the test takers see CEFR 2001, Sections 4.1 and 4.3.

The CEFR Companion volume presents new scales with descriptors for online interaction and mediation activities according to the four domains, situations and roles (Appendix 5).

If the assessment is part of an educational cycle, alignment between curriculum, delivery and assessment is a primary objective. The British Council sets this out as the comprehensive learning system (CLS). The three core elements need to function as a system and be rooted in a single philosophy of learning supported by clearly defined models of language ability and progression and a measurement model. The system is related to a specific educational and social context and needs to respond to the expectations of the stakeholders.

For further information on CLS see *Aligning Language Education with the CEFR: a handbook* (O'Sullivan, 2020).

2.4.2.2 Competences, strategies, activities and tasks

Authenticity

One of the most important features of the test tasks is authenticity. Authenticity reflects the relationship between test content and target language use (TLU), the tasks people use language for in the real world. Authenticity is a matter of degree. A higher degree of authenticity is normally a primary objective for a language test and contributes significantly to test usefulness (See Section 2.4.5). Test developers need to analyse the tasks in the TLU in order to try and replicate them in the testing situation. At the same time, they need to analyse the mental processes involved in the interaction between the test taker and the task in the TLU and replicate them in the test task.

In Bachman and Palmer's terms, situational authenticity represents the perceived similarity between the characteristics of the test task and those of the task in the TLU, while interactional authenticity refers to the naturalness of the interaction between the task and the test taker and the mental processes which accompany it. Authenticity is a challenging characteristic to obtain for a test task, with many factors which might hinder it and which test developers should try to avoid, e.g. artificial test design, unrealistic or extremely complex language used in rubrics, or an inadequate physical or social context of the test, above the artificiality of the testing situation as such. Variables related to individual test takers, whose perception of the authenticity of the task may differ greatly, need to

be considered. Test developers might find different solutions for increasing the degree of authenticity of the test tasks. For example, consider a multiple-choice item requiring the test taker to listen for specific information in a recording presenting the weather for different weekdays, and select a day of the week in answer to the question 'Which day will it be sunny?' This item can be made more situationally authentic if an everyday context familiar to test takers, such as a radio weather forecast, is created. Multiple-choice items are notorious for not reflecting real-life cognitive processes, but this item can be made more interactionally authentic if the test taker is given a reason for listening which is close to their real daily life and if selecting the response requires similar cognitive processes. For example, for test takers for whom picnicking is a familiar activity, the question could be preceded by 'you are planning a picnic that week and must select a suitable day' because this will turn the listener's attention to the relevant signals in the text in the same way as would be the case in a real-life context, and the real-life task is, like the test item, a selection among several possibilities. If a real-life task is too complex to be integrated in a test, it may still be possible to include components which may not stand alone in real life (thus reducing situational authenticity) but which are interactionally authentic because they tap competences which are essential to the real-life task (Weir, 2005, p.148).

In language tests, we often have to balance different aspects of authenticity to create an appropriate task. For example, it may be necessary to adapt materials and activities to language users' current level of language proficiency in the target language. This tailoring means that while the materials may not be pulled from real-life sources, or even linguistically authentic as typical representations of the language tasks in the real world, the situations that language users engage with, and their interaction with the texts and with each other, *can* be authentic. Another example of balancing authenticity with other considerations is the interactivity of tasks: in many cases, it is impractical to have test takers engage in real-time interaction, either because of the demands of large-scale rating, or because of concerns about consistency of the testing experience. A test of oral production, then, may involve a test taker speaking into a microphone with either no interlocutor at all, or an artificial intelligence (AI) avatar. Such a test may not replicate an authentic communicative situation, but may be able to capture underlying skills and abilities used in such situations.

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. Using linguistic corpora can be helpful here, if there is a corpus of relevant language use available: a corpus can provide guidance about the relative frequency of vocabulary or grammatical structures, or the usual uses of idiomatic expressions.

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
- Analysing the mental processes involved in carrying out the real-life task and replicating them in a test task

See also ALTE 2020, Section 2.3.1.

Integrated assessment

As mentioned in Chapter 1 (1.3.2.2), language users typically engage multiple competences and strategies to complete linguistic tasks, for example, speaking and listening in a conversation. The CEFR framework as updated in the CEFR Companion volume acknowledges this in the model by including scales for interaction and mediation. The natural integration of competences poses a challenge for testing. The traditional approach to this challenge is to try to isolate one primary factor and reduce 'interference' from other factors, such as when tasks testing listening comprehension are designed to have any written instructions or questions use the simplest possible language so as to eliminate any variance in reading comprehension from the measurement. Another approach is to design tasks that deliberately focus on multiple competences, known as integrated assessment. In some cases, integrated tasks are particularly appropriate due to the similarity to tasks in real life

(e.g. in the context of testing for admission to higher education, listening to a text, possibly an excerpt from a lecture, and taking notes, or reading a text and summarising it as part of an oral presentation). Determining how to score and assess responses to tasks where more than one competence is involved may pose challenges for measurement, as will be seen in Chapter 3, Section 3.2.2.

See Chapter 2, Section 2.4.3 and Chapter 3, Section 3.2.2 for more information.

Plurilingual competence

While learning and being assessed into a new language, the language learner/test taker builds on and along with the competences in their first language(s) and culture(s) and becomes a plurilingual speaker in an intercultural environment. The test developer might not want to keep the tasks in the language test in complete isolation from the ability of the test taker to use another language or other languages. For example, tasks might provide rubrics in one language and require the response in a different language or in two different languages, especially when the learning environment is homogeneous. The rating instruments might be designed to reward the test taker's capacity of drawing on their plurilingual resources in their responses instead of penalising this.

See CEFR 2001, Chapter 6, CEFR Companion volume, Chapter 4, ALTE 2020, Introduction for more information. Chapter 7 in CEFR 2001, on tasks and their role in language teaching, has implications for how tasks might be used in testing.

Chapters 4 and 5 can be consulted for information on: the focus of tasks, e.g. showing detailed comprehension of a text, etc. (sections 4.4 and 4.5); what is to be tested, e.g. competences and strategies (Chapter 5), along with Chapter 2 in the CEFR Companion volume.

2.4.2.3 Texts and prompts. Sources and types

The definition of situational authenticity has often been taken to be that the text was created by users of the language for communicative (not pedagogical) purposes (authenticity is used in this sense in the CEFR 2001). However, a better term for this concept is *genuineness*. Genuineness is an attribute of the text itself – either a text was created by language users for communicative purposes, or it wasn't – and is an absolute property, unlike authenticity, which is a matter of degree. Authenticity is a characteristic of the relationship between the text, the purpose for which it was originally intended, and the reader or listener (Lewkowicz, 2000; Widdowson, 1978), in that it involves the plausibility, not necessarily genuineness, of the setting to the test taker. The use of a text purpose-written for a test to showcase specific grammatical and vocabulary specifications, for example the weather report example above, could be authentic, even though the text itself would not be genuine. Similarly, the use of a passage from a novel as the basis for a task asking test takers to identify which words are verbs might not be very authentic, even though the text itself is genuine. A text generated by AI such as a Large Language Model (LLM) could well be genuine, as there are many situations in real life in which language users use AI to generate language for communicative purposes, for example news articles. The genuineness of the AI-generated text, like the genuineness of human-generated text, will depend entirely on the purpose for which the material was generated (real-life communication vs. didactic purposes). Further, the use of such a text would be just as authentic as a human-generated text if the context of use is relevant to the test taker and is sufficiently representative of the universe of texts that might be used in this context, for example in a task asking test takers to read an online article for the gist. The choice of using AI-generated or human-generated texts, then, should generally be based on interactional and situational authenticity, and not solely on what kind of 'language user' generated the text.

Tasks focusing on oral and reading comprehension are usually based on recordings and text fragments of various lengths, while tasks focusing on spoken and written production and interaction are usually based on different types of prompt (e.g. input texts, images, recordings and videos). To keep the test content as close as possible to TLU, test developers may wish to use as input recorded or written genuine texts (i.e. texts which were created for real-world use, rather than for pedagogical purposes). In this case, input recordings could come from radio broadcasts, podcasts, social media posts, public announcements, lectures, etc., while written texts could come from magazines, websites, social media posts, blogs, brochures and flyers, etc. Sometimes the input material needs to be adapted, for various reasons, such as matching length requirements, eliminating elements of vocabulary, grammar, and syntax which might exceed the level being tested, and removing references to themes that could be deemed inappropriate for certain test takers (e.g. children).

The prompts can also be taken from real-world contexts or generated with the help of AI. For example, a spoken production task on the topic of 'Professions' could use a series of photos as input. These photos could be created by the test developers, sourced from various websites, or produced using AI.

The specifications should provide guidance about whether texts must be genuine with no modifications, whether they must be genuine but can be modified, or whether they can be purpose-written for the test. If texts must be genuine but can be modified, specify the types of acceptable modifications (for example, whether audio inputs can be re-recorded).

Materials produced with the help of AI can be considered genuine, given that content is frequently generated with AI assistance in real-world situations (e.g. news articles, studies, promotional materials, announcements, etc.). Just as with human-generated material, input materials and prompts produced with the help of AI need to be verified and confirmed by either the person responsible with the collection of input texts and prompts or the item writers. Test developers should consult the test sponsor and stakeholders regarding any requirements or prohibitions concerning the use of AI in producing input materials and prompts.

Using genuine input material and prompts does not guarantee that the tasks are authentic, and tasks can be quite authentic even with material that is purpose-written. Genuineness is a property of the origin of the material, whereas authenticity is about its use.

For text sources see the CEFR 2001, Sections 4.1 and 4.6. Please be advised that the CEFR 2001 uses the term authenticity to refer to both the input material and the tasks, especially through Chapter 4 (Section 4.1) and Chapter 6 (Section 6.4.3.2). Authentic written or spoken texts are defined as 'produced for communicative purposes with no language teaching intent' (CEFR 2001, Section 6.4.3.2). The present Manual names this kind of material genuine and the tasks authentic.

For text types used as input see CEFR 2001, Section 4.6.

For types of prompts used in tests of oral production and interaction see CEFR 2001, Sections 4.3 and 4.4.

2.4.3 Test format and practical considerations

The test provider must also determine a number of practical 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 models may be consulted (see Section 2.8). After the try-out phase and use in a live context, the timing may be revised (see Section 3.3.3). In digital adaptive tests the time allotted to different test takers in the same test (even in the same test session) might not be equal, the same as in the case of the number of items. The test will be stopped when the test taker's ability is considered proven.
- **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. Bear in mind that if you do embedded pretesting (see Chapter 3, Section 3.4.3), not all the items will count for scoring, so you will need a longer test than if all items are counted. If you do adaptive testing, you may be able to have a shorter test because most of the items will be tailored to the specific ability levels.
- **The number of items per component**. If the test aims to measure reliably different aspects of ability, then a sufficient number of items per component will be needed. When data is available (e.g. after try-out), reliability can be estimated to assess the appropriateness of the number of items considered (see Appendix III).
- **The item types**. Items can elicit selected or constructed responses. Selected response types include, for example, multiple choice, matching or ordering. Constructed response types include

short responses or extended writing. Item types have different advantages and disadvantages. See ALTE (2005, pp.111–134) for more information on item types.

- **How responses are recorded**. In a paper-based test, the type of response could be inserting letters in boxes, circling the correct option, writing short or long answers, matching (by drawing lines). An answer sheet might be used. In a digital test the answers could be given by drag and drop, clicking, writing short or long answers, etc. For the speaking part, determine whether you need to record the responses. This might be useful for asynchronous rating, rater monitoring or to process an appeal request.
- **How test takers interact with the audio input**. In real life, people hear spoken language only once and must process it on the fly; however, they may also have the opportunity to ask for repetition, and they will typically have considerable information about the context. A balance between authenticity and fairness might need to be considered. Determine whether audio will play only once, if it will play twice or the test takers will be able to replay or control the start of the audio and whether contextual information (an orientation to the situation, or the tasks themselves) will be presented orally or in written form.
- **The length of texts**. The total and individual length of input texts, e.g. as measured in words, for the receptive components (e.g. oral comprehension and reading comprehension) needs to be decided. Models can provide guidance about how long is practical (see Section 2.8). The number of required words for constructed responses needs to be indicated in the specifications and in the task rubrics.
- **The grouping of items**. Items can be discrete or grouped in relation to an input text. A set of discrete items consists of short items which are unrelated to each other (e.g. items based on sentence-length input verifying grammar knowledge). There are also tasks in which a number of items are grouped together, e.g. relating to a reading or listening text – binary choice, multiple choice, short answers, filling out blanks, etc. They can use longer stimuli that better reflect authentic language activities and are considered generally more appropriate for communicative language testing. Tasks with discrete items and tasks which group items in relation with a text can be combined in tests.

See ALTE (2005, pp.135–147) for more information on task types.

2.4.4 Considerations on marking and rating

In relation with marking and rating the following need to be considered:

- **The number of marks to give each item/task/component**. The items, tasks or components might be weighted differently; 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 – e.g. the perceived difficulty of the knowledge or ability tested through the item; the use of the test – e.g. in an occupational test for hotel receptionists the oral skills might be weighted more than the written ones due to the profile of the main tasks in the job.
- **Setting the cut-off scores**. The cut-off score represents the minimum score a test taker needs to achieve in order to pass the examination or to meet a predefined standard or a particular proficiency level (e.g. CEFR B1). It needs to be determined how the cut-off scores will be set. The cut-off score is usually the result of a standard setting process which is conducted by the test developer. Standard setting helps answer the question: *at what point on the score scale of a particular test can we conclude a test taker has sufficiently demonstrated the proficiency described for a particular level within a set of standards, qualification levels or framework of proficiency*?' (*Aligning Language Education with the CEFR*, p. 46). There are several methodologies that have been developed for aligning tests and assessments with standards, qualification levels and frameworks defining criterial features of proficiency; at this stage, it is not necessary to specify all the details of the methodology, but having a general plan may help determine features of rating scales and mark schemes, particularly regarding how items will be weighted (see above) and what the rating scales will be like (see below). See Section 5.3 for more on standard setting.

- **The characteristics of rating scales**. Will task-specific scales be used, how long will each scale be, will scales be analytic or holistic? (See the more detailed discussion on rating scales below.)
- **The way items are scored/marked**. Will humans, computers, or AI marking instruments be used? Will there be a combination of human/automated marking? (See Sections 5.1 and 5.2)
- **How many ratings will be collected?** How will the final rating be arrived at (e.g. consensus, senior rater decides, additional rating in case of disagreement)? (see Section 5.2)
- **The possible uneven profile of the test taker**. Typically, standard setting is conducted on one component of a test, for example a written interaction component, so that a test taker might achieve different levels on different components (see CEFR Companion volume, p. 39). Best practice is to communicate a grade (e.g. number of points, percentage, placement on a scale) for each component separately to reflect the proficiency profile of the test taker. Some stakeholders, however, require an overall grade or level. If an overall grade or level is needed, how will it be calculated from the component grades? Will there be a compensatory procedure (i.e. allowing strengths in one skill to counterbalance weaknesses in another when calculating the total score for the test)? See ALTE (2023) for a discussion of this issue and Section 5.4 in this Manual for reporting considerations.
- **Partial qualifications**. Will the test provider or the test sponsor include 'partial' qualifications as an option for the test takers, 'appropriate when only a more restricted knowledge of a language is required (e.g. for understanding rather than speaking)' (CEFR 2001, p.1)? Certification, in these cases, will be partial, 'taking responsibility only for certain activities and skills' (CEFR 2001, p.6).

Rating scales

Whether the test takers' written or spoken production and interaction are rated by human raters, by automated systems or a combination of the two (e.g. one human rater, one automated system for each response), most approaches to testing proficiency depend on some kind of rating scale.

A rating scale is a set of descriptors which describe performances in different bands, showing which mark or grade the corresponding performances should receive. Developing the rating scales, possibly with the contribution of AI, is an essential element of operationalising the construct, as it is through the rating scales that the desired characteristics being measured are articulated. The quality of measurement of productive activities and strategies, particularly, depends on the reliability and accuracy of the work done by raters (human or automated systems) and the quality of the scales. Rating scales reduce the variation inherent in the subjectivity of human judgements and guide the assessment performed by AI systems. After rating scales are built, they can be tested empirically (e.g. in the try-out phase) to detect possible problems and can be refined for better performance and reliability. (For a discussion of human and automated rater training, see Section 5.2).

To build valid and reliable rating scales, various aspects must be taken into account, chiefly the construct of the test (See Section 2.4.1), how the scales will be used, and a series of practical concerns. 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. They are sometimes called grids (CEFR 2001, 9.2.2.2). All approaches may relate to a similar language proficiency construct described in similar words – the difference is in the judgement the rater or the automated system is required to make. See Weigle (2002) for further information about writing scales and Taylor (2011) for information about scales for speaking.
- **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 (CEFR Companion volume, Appendix 4).

- **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 is preferable, and the CEFR descriptor scales are a useful source for constructing such rating scales.

You also need to consider the following:

- **The identification of rating criteria** (how many and which ones, for example task completion, coherence, grammar, etc.).
- **The number of bands and how these will be transformed into scores** (cf. 9.2.2.2 in CEFR 2001).
- **The definition of clear descriptors for the different levels of competence**, which may be inspired from the CEFR 2001 and the CEFR Companion volume, from any other relevant framework or curriculum or created according to the specific needs of the programme.
- **The formulation of the descriptors following specific criteria** (e.g. positiveness, definiteness, clarity, brevity, independence – CEFR 2001, Appendix A).
- An alternative or additional approach to scales might be **checklists** (giving marks based on a list of yes/no judgements as to whether a performance fulfills specific requirements or not) (Weigle, 2002).
- **Comparative judgement**, commonly associated with assessment of written productions, can also be used for rating and for training human raters and automated systems (CEFR 2001, Appendix A).

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

- All rating depends on the raters understanding the proficiency levels and the descriptors in each band of the rating instrument
- Models are essential to defining and communicating this understanding
- The test tasks used to generate the rated performance need to reflect in the descriptors of the scales

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 or other frameworks' approach of describing performance levels in recognisable ways, using positive and definite statements. This does not change the fact that performance samples (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. Large numbers of models are also crucial in training AI systems for accurate use of the rating scales in scoring written and oral responses.

The CEFR promotes thinking and working in terms of criterion proficiency levels (the six main levels ranging from A1 to C2, with their main features captured in specific descriptors). Pre-A1 was added, and the plus levels and C level descriptors were refined in the CEFR Companion volume. There are two aspects to defining levels: what people can do and how well they can do it. In an exam, the what is defined through the tasks which are specified. If and how well these tasks are done is what raters have to judge.

This 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 are not always explicitly referred to in defining what a 'passing' performance means (see *Aligning Language Education with the CEFR*).

CEFR 2001 (p.188) discusses subjective assessment. Building a common understanding of the construct and reliable assessment instruments and training raters adequately, as well as regular calibrating, are all ways in which subjectivity in assessment can be reduced.

The design phase thus ends with documentation of initial decisions concerning the purposes the test is to serve, the skills and content areas to be tested, and details of the technical implementation. Decisions are also taken in this phase about how the tasks are to be marked, how rating scales

should be developed, how tests should be administered (Chapter 4), and how markers and raters or marking and rating instruments should be trained and monitored (Section 5.2).

2.4.5 How to balance test requirements with practical considerations

At this stage of specifications development, the proposed test design must be balanced against practical constraints. Information about constraints is gathered in the planning phase, together with test requirements and background information (Section 2.3; see also Section 1.5.3 and Chapter 4). The test developer must reconcile requirements and constraints, and this must be agreed by the test sponsor. Bachman and Palmer (1996, Chapter 2) provide a framework to do this through their concept of test usefulness. The ALTE *Principles of Good Practice* (2020) also has a section on test usefulness. From these two works we extract the following eight 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 (e.g. financial, technical, personnel).
- **Fairness and justice** – The content, scoring, and administration of the test treat all test takers equitably, and the values implicit in the test construct and the social impact of the test are just (See Section 1.5).
- **Quality of service** – Provision of secure examination materials, confidentiality of examination data and results, procedures to handle enquiries about results and appeals procedures, provision of documents on the use of test scores for stakeholders.

These qualities tend to compete with each other: for example, increasing task authenticity may lead to a decrease in reliability – e.g. a speaking task normally has a higher degree of authenticity in comparison with a binary choice task given the fact that it mirrors communicative situations in the real world; on the other hand, the speaking task might have a lower degree of reliability than a binary choice task because the measurement is more subjective, possibly leading to variability and reduced consistency in the score; the generalisability of the results of a speaking task might also be more limited in relation to the situations in the real world, especially when the test is rather general, not for specific purposes. For this reason, it is the effort to find the best balance between these qualities which is important in increasing test usefulness overall.

2.4.6 The draft specifications

The draft specifications will be written based on the information gathered in the planning phase (see the example in Table 1) and on the construct, having in view all the considerations on content, form, administration, and marking guidelines and instruments documented during the design phase (discussed in the previous sections).

The specifications resulting from the design phase are in a draft shape. After the try-out (see Section 2.5), these specifications will be brought to their final form. After amendments a new try-out phase might be necessary. If the test is high-stakes, specifications are particularly important because they are a fundamental tool to ensure the quality of the test and, where published, to show others that the recommended interpretations of the test results are valid.

Specifications also help to ensure that test versions have the same basis and that the test correctly relates to a teaching syllabus, or other features of the testing context, so they are important not only for high-stakes, but also for low-stakes tests.

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 Section 2.8) and can serve as a useful starting point. The specifications should refer to: content (themes, type of input texts and text sources), tasks and response (targeted activities, type of response required, number and type of tasks), marking and rating instruments, mode of delivery, etc.

The writing component of the fictitious B1 test for the purpose of obtaining citizenship (Section 2.4.1) might have two tasks, one representing written interaction and the other one written production. Taking further the construct for written interaction presented in Section 2.4.1, the corresponding content and task specifications for the part of written interaction might look like Table 2 below.

Table 2 ***Example of content and task section of specifications, Level B1: Written interaction***

Targeted activities	Type of answer	Topic areas	Characteristics of task
– writing and properly organising informal texts of correspondence (email, personal letter) related to personal and professional data and experiences (e.g. daily activities, preferences, hobbies, future projects, familiar people and places, professional domain, education) – expressing personal impressions, attitudes and reactions to received information and/or personal experiences – expressing and simple argumentation of opinions related to general topics (e.g. visits, travelling, public services, movies, shopping) – asking for and offering information, suggestions, recommendations and advice – highlighting main points in message related to concrete themes – adequate use of grammar and vocabulary elements; connecting sentences through proper, simple linkers, clearly indicating chronology of events	– informal correspondence (emails, letters, posts): describing experiences in the country of resettlement, recommendations, suggestions, advice, apologies, thanking, simple argumentation	– life in town/countryside – shopping – health – education – media and communication – relationships – jobs – public services – travelling – cultural activities – sports and hobbies	– authentic task – test taker production will be a response to a received message – familiar topics, of general interest, related to daily life, school or work – adequate register of the input message (familiar) – prompts to support the test taker in constructing the answer – number of words to be written is indicated

2.5 Try-out and final specifications

The aim of this phase is to 'road test' the draft specifications and the rating instruments which were proposed and to 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 Chapter 3 of this Manual. Information about these materials can be gathered in a number of ways:

- Trying out the sample items (asking some test takers to sit the test) and the analysis of the responses and possibly of collected feedback (see Section 3.3.3 and Appendix II)

- Consulting colleagues
- Consulting other stakeholders

The try-out should ideally be conducted using test takers who are the same or similar to the target test takers. The pretest or trialling test should be administered under exam conditions, as the live test would be. However, particularly for very preliminary versions, piloting with colleagues or a small number of test takers will still be useful even if it is not possible to replicate a live test administration (e.g. perhaps there is insufficient time to administer the entire test), or if numbers of test takers are small. It can provide information about the timing allowance needed for individual tasks, the clarity of rubrics, appropriate layout for the response, etc. For oral components, observation of the performance (e.g. via a recording) is highly recommended. In addition, verbal (think-aloud) protocols and other methods of investigating test taker cognitive processes can be useful as a component of piloting.

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

Information from piloting, pretesting, and/or trialling also allows fairly comprehensive mark schemes to be devised and rating scales which were produced in the design phase to be tested/checked (see Section 2.4.4 for features of rating scales). Features can be identified in test taker performances which exemplify the ability levels best. Rating scales can be revised with input from raters and from other stakeholders.

Input from raters. The scales should be used by a number of raters who should provide comments, e.g. on their clarity, ease of use and consistency. The results should be analysed, qualitatively and/or quantitatively (see Appendices II and III).

Input from stakeholders. The stakeholders can be consulted in a two-step process: first when the relevant descriptor scales are established and second when realistic goals for each one are determined. Further piloting, trying out and revision of scales may be necessary (see Section 2.6.1). For the qualitative validation, ideally, the scales should be presented to representatives of all groups of stakeholders, and their feedback analysed. It is important to remember that such feedback may help us improve certain inconsistencies of the draft, but that the overall shape of scales should not be questioned or, otherwise, we would end up in a vicious circle of revision and rebuilding scales. The foundation of our scales are the technical decisions derived from the construct of the test.

There are different and very sophisticated ways of validating rating scales quantitatively (Knoch, 2011), which time and economic constraints may make impracticable. Other methods, more attainable, can give us accurate psychometric insights into the internal functioning of our scales. Appendix III illustrates a way to check the appropriateness of ratings scales, based on Linacre's work (2009) and the use of Many-facet Rasch Measurement (MFRM).

Other kinds of research may be needed to resolve questions which have arisen during the try-out phase. It may be possible to use data from the try-out phase 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)?
- Do items and tasks really test the competence area they are supposed to? Statistical techniques can be used to determine this (see Bachman, 2004). Qualitative methods can also be used, especially for smaller-scale programmes (e.g. verbal protocols).
- 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 as the basis for live tests.

Once the content of the specifications has been agreed upon, it can be used as the basis for a set of documents directed at different stakeholders. They can vary in content and format according to the focus specific to their users. For example, detailed specifications can be produced for item writers, with information other stakeholders would not need to know or would find difficult to understand (see Section 3.1.2). The test takers and the personnel involved in the administration and management of the tests are other groups of stakeholders for which specifications can be presented in an accessible and tailored way.

2.6 Relating with stakeholders

2.6.1 Cooperating with stakeholders

Cooperation with stakeholders needs to be constant throughout the specification development stage. As indicated under Section 2.3, cooperation with different groups of stakeholders in the phase of planning is vital for collecting relevant information for test design. In the design phase the stakeholders are consulted in relation to the initial specifications and in the process of design of assessment instruments. Stakeholders should be involved in reviewing the specifications in detail, enabling proper evaluation of them. The stakeholders with whom test developers cooperate in the try-out phase might be potential test takers, language teachers and score users (e.g. higher education recruiters, employers, state agencies, migration authorities, etc.). Analysis of results from potential test takers and feedback from teachers and assessors will help improve the tasks and the assessment instruments. Feedback from other groups of stakeholders will inform on task adequacy to the purpose.

2.6.2 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.

At this stage of the Testing Cycle, communication with test takers and other stakeholders should be considered with regard to the following:

- The estimated number of hours of study expected as preparation for the test.
- How sample tests will be made available.
- How scores will be reported: will each component or skill be reported separately, or will there be an overall score instead of or in addition to component scores? If there is an overall score derived from component scores, how is the overall score calculated: by averaging all components equally, or by weighting some of them more than others? What justifies the approach chosen? Do different stakeholders have different requirements for this?
- What information will be given to potential score users both before and after the test.
- How the test will fit into the current system in terms of curriculum objectives, classroom practice, migration requirements, professional qualification requirements, etc.
- What the stakeholders will expect the test to be like.
- How the test takers' personal data will be protected (See Sections 1.5.1.5 and 4.2.).
- What the correct/just use of test results is and how test developers and providers may prevent the use of test results to violate the best interest of the test takers (e.g. this test was not developed for migration purposes).

- Whether AI was used/is anticipated to be used or not and how it was/will be used in any stage of test development, administration or marking.

If significant revision of the test is performed, the stakeholders need to be informed in advance of the changes and their potential consequences (e.g. transitioning from paper-based to digital administration, changes in the format of the test, fewer or more tasks and consequential changes in the time allotted for response, changes in the potential use of the examination, etc.). See Chapter 6 for information on test revision.

Chapter 4 of the CEFR 2001 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.

2.6.3 Additional materials

In addition to test specifications, stakeholders will find it very useful to see full sample materials (see Chapter 3 for more information about producing materials). The sample materials can include test versions and examples of answer sheets, and also audio or video recordings used in tasks that involve listening. 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 productive tasks, it may be useful for test takers to see sample performances. These can be prepared if the materials have been trialled, or used as a live test in the past. In both cases, make sure that you have the consent of the test takers having produced the material to put them to this use (See 1.5.1.5). In addition, advice to test takers can be provided to help them prepare for the test, both for productive and for receptive tasks.

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 or that a test should be revised and why – what can they tell the test developer about its purpose and use?
- What will the educational and social impact of the test be?
- What competence(s) and what level of language performance need to be assessed?
- What type of test tasks are appropriate to achieve this?
- What practical resources are available? (e.g. premises, personnel, etc.)
- What technical resources/means are necessary/available?
- Who should be involved in developing test specifications and sample test materials? (e.g. in terms of expertise, providing information, etc.)
- How will the content, technical and procedural details of the test, administration, and scoring be described in the specifications?
- What sort of information about the test needs to be given to users and groups of stakeholders? (e.g. a publicly available version of the test specifications, and how this should be distributed, use of test results)
- How can the test be tried out?
- How can stakeholders be consulted?
- How can stakeholders be best informed about the test? What information do they need in order to select an appropriate examination?

- How will the test developer make sure that the examination is as fair as possible for test takers of different backgrounds (e.g. gender, age, educational background, etc.) and how are these communicated to test takers and other stakeholders?

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 and Federation of Migros Cooperatives (2004), University of Cambridge ESOL Examinations (2004), CIEP and Eurocentres (2005), Bolton et al. (2008), Grego Bolli (2008), Council of Europe and CIEP (2009), CIEP (2009), *Aligning Language Education with the CEFR: a handbook*, ALTE (2020), ALTE (2023).

Sample test specification formats may be found in Bachman and Palmer (1996, pp.253–334), Alderson et al. (1995, pp.14–17) and Davidson and Lynch (2002, pp.20–32).

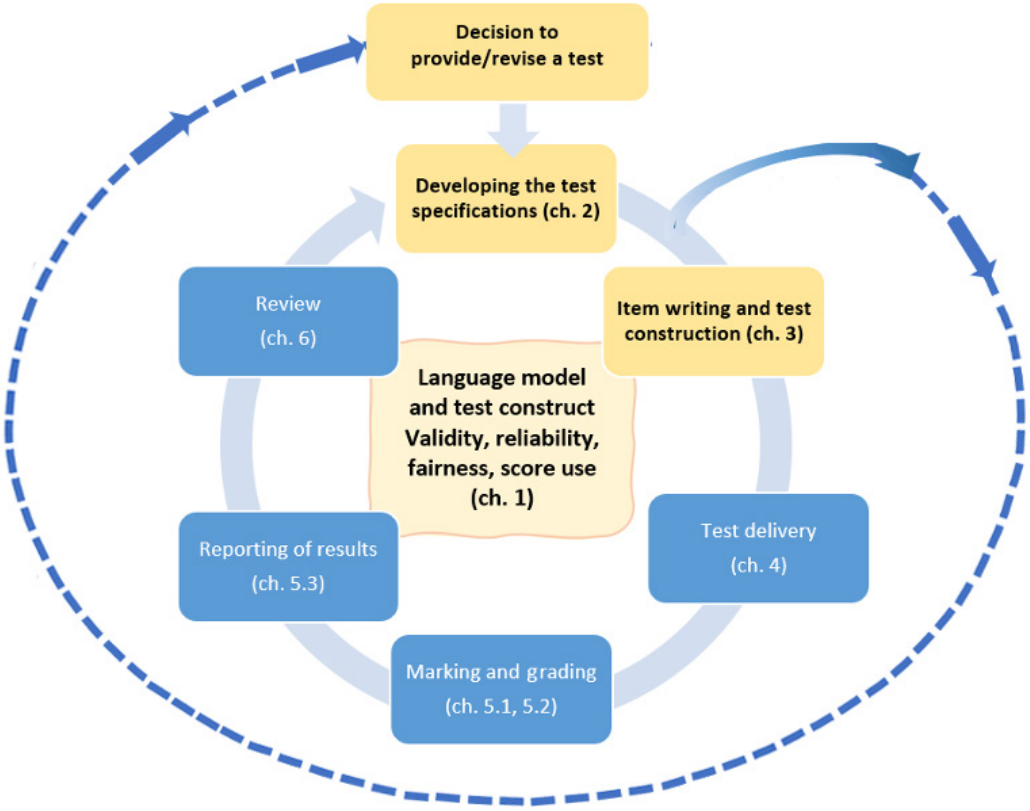
A number of grids are available which provide templates for describing and comparing tasks, including ALTE (2014a,b,c), Figueras et al. (2005), Knoch (2011).

Next steps

Chapter 2 presented the stages which stem in the decision to provide or revise a test: planning, which is concerned with gathering information on the characteristics of the test takers, purpose of the test, etc.; and design, which shows how the construct and specifications are built. The relation with the stakeholders was analysed and presented in connection with every phase in the stage.

Chapter 3 will show how test developers can build on the construct and specifications and proceed with test construction. The process of item writing and content management will be shown in all its dimensions: recruiting and training item writers, steps of item writing, and quality control and construction of test versions.

3 Item writing, test construction and content management



This chapter offers a structured approach to the production and management of test content, providing practical suggestions, guidelines, and ideas that can be useful for both experienced institutions and practitioners and those new to the field. While the chapter presents a range of good practices, it is important to emphasise that these are non-exhaustive and should not be seen as prescriptive. Institutions are encouraged to adapt and customise the content of the chapter based on their specific needs, resources, and goals.

Effective language assessments hinge on meticulous test development. This process involves several stages, each contributing to the creation of high-quality test materials.

The process of test construction starts with completed test specifications (see 2.5) and breaks down into distinct yet interconnected stages:

- Producing content
- Quality control
- Constructing test versions

It also requires considerable preliminary work before the stages can begin, discussed in the next section.

3.1 Preliminary steps

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

- Item writer recruitment and training
- Materials management

3.1.1 Item writer recruitment and training

The term 'item writer' refers to a professional who is responsible for preparing test content, such as input texts, items, or tasks. Depending on their expertise and the needs of the test provider, item writers may be internal staff or externally recruited staff, and each item writer may be assigned a single item or task type, a single test component or an entire test version.

Item writing is a process that demands both creativity and discipline, as writers must adhere to test specifications and/or assessment frameworks of reference (for example national or local language curriculum frameworks). Thus, item writers play a crucial role in the test development process, ensuring that the content is aligned with the test's objectives and specifications and is fit for purpose.

Item writers should have a background in language teaching, or in creative or subject-specific writing. Understanding of language learning and the language itself, experience in preparing students for similar tests, or experience in marking or oral examining may be particularly useful.

When selecting item writers, the test provider must determine the minimum professional requirements they should meet. These requirements may include:

- A specified level of language proficiency
- Familiarity with the testing context
- Familiarity with the CEFR
- Teaching experience
- Experience in preparing students for similar tests
- Experience in oral examining or marking

While knowledge of existing tests and assessment principles is beneficial, these can be addressed through training (see ALTE, 2005) and do not necessarily need to be part of the minimum requirements. Ongoing training, monitoring, and evaluation are crucial for the continued professional development of item writers.

After recruitment, item writers should receive training that grounds their work in the CEFR (see Sections 1.3.2 and 2.4.1 for the relevance of CEFR to the construct) and in the CEFR linking process as outlined in the *Manual for Relating Language Examinations to the CEFR* and *Aligning Language Education with the CEFR. A handbook*. Training should begin with **Familiarisation**, ensuring writers fully understand CEFR levels, descriptors, and illustrative scales. It should then introduce the other stages — **Specification**, **Standardisation**, **Standard Setting**, and **Validation** — so that writers understand how their items contribute to the overall alignment of the test. Embedding these stages into training helps ensure consistency, level-appropriateness, and conceptual alignment throughout the item development cycle (see *Manual for Relating Language Examinations to the CEFR* and *Aligning Language Education with the CEFR*). Item writers will also need to be provided with the relevant specifications and submission requirements, and, depending on the complexity of these, may need training in how to use them (see next section).

3.1.2 Item-writing specifications and submission requirements

Item writers may be the same people who are 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. However, internal staff may not always have the level of familiarity with the specification details required to write effectively, so some training will likely be needed. All item writers, regardless of how much training they need, should have access to all the materials they need to ensure the effectiveness and consistency of the item writing. The test specifications, as well as details about how the items should be submitted, should be clearly communicated to item writers (see ALTE, 2005). Test specifications may be tailored specifically for item writers, but they must reflect the *full scope* of the specification development process outlined in Chapter 2, rather than focusing solely on test content. As described in Chapter 2, final specifications encompass not only the themes, text types, and tasks to be included, but also the construct definition, the mental and cognitive processes

expected from test takers, and the marking and rating evaluation criteria and procedures through which productive tasks will be evaluated. These elements — developed through the planning, design and try-out phases — provide the conceptual foundation needed for item writers to understand how items operationalise the construct and what types of performances they should elicit. Even when working with a small team, it is therefore essential to provide item writers with specifications that go beyond content points and address all facets of the test in alignment with the processes presented in Chapter 2.

3.1.2.1 Item writing specifications

The specifications serve as the bridge between the test construct and the practical work of item writers. During the try-out phase described in Section 2.5, a lack of consistency in items produced according to the draft specifications may lead to revisions to ensure that the specifications cover all necessary elements. As mentioned above, test specifications may have different versions with different levels of detail; versions intended for item writers may contain more detail on copyright, sources, and word counts, for example, than versions intended for score users. Some elements of the specifications that are particularly useful for item writers include the following:

- **Item type**. The item formats, such as multiple choice, short answer, open-ended questions; tasks or prompts for written and spoken production.
- **Construct alignment**. How each item/task type is linked to the construct elements defined in Chapter 2, with the communicative activities, strategies, and competences targeted by the item. If the test is aligned to the CEFR, the CEFR level(s), scales and illustrative descriptors are specified for each item type and task; the level of the input text (for text-based items such as those testing oral or written comprehension), rubrics, and item texts may also be specified. Specifications of the mental processes expected from the test taker (e.g. inferencing, scanning, organising information, producing connected discourse) are also helpful. For integrated tasks, item writers need to know which competences are central, how they relate to the construct, and how they will be weighted in marking.
- **Language limitations**. Specifications for word counts or text length, and any language use or vocabulary lists.
- **Topic restrictions or sensitivities**. Taboo lists of topics which are not appropriate for some groups of test takers (e.g. young learners) or for all test takers (e.g. religion, drug abuse, violence).
- **Text and prompt characteristics**. Specifications about whether texts must be genuine (taken from materials used in real-life situations), or whether they can be purpose-written, and if they can be purpose-written, whether they must be written by humans or whether AI may be used; specifications on the genre, register or style required.
- **Copyright considerations**. Specifications for whether copyrighted material must, may, or must not be used in the test (e.g. photos, input texts, audio or video recordings).

3.1.2.2 Submission guidelines

In addition to the test specifications detailed for item writing, item writers will need procedural documents so that they know what to submit. For example:

- **Text source**. Depending on the specifications for copyright considerations, item writers may be required to provide citation information. Depending on the specifications for use of AI, item writers may need to provide information about how texts or images were generated.
- **Approval process**. Include procedures for approval of drafts of texts, items or both before item development continues. State whether texts should be submitted for approval before item development.
- **Guidelines for quality**. Provide expectations about the degree of the background knowledge expected of test takers, or reminders about avoiding unnecessarily complex rubrics that obscure the intended focus.
- **Submission requirements**. Indicate what item writers must submit, for example the task-rubrics, input texts or scripts, media (images, audio, video or prompt descriptors).

- **Expectations for audio and visual materials**. For visual prompts, clarify whether a specific prompt is needed, if item writers should find/produce visuals themselves (e.g. through AI or other sources) or provide a detailed description of the desired visual prompt. For audio or video materials that are not taken directly from real-life sources, indicate whether the submission is expected to be just a script, the final version of the audio or video material or whether the testing organisation will produce a final audio/visual version after the review process has made changes to the input text.

Along with the specifications and submission guidelines, it is important to include sample materials. The test specifications will typically have exemplar items, but it may be necessary to have additional materials, for example, items with specific problems to avoid, or, for multi-level tests, examples at each targeted level, if the specifications contain just a single example per item type.

3.1.3 Managing materials

The test provider must set up a system to collect, store, and process the items. All test materials should go through a standardised quality assurance process, such as editing, piloting, and reviewing, though the particular process may be different depending on the item type. For this reason, it is necessary to track the stage of any item 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 might include:

- A unique identification number for each item
- A record of the stages completed, changes and other information, including metadata tags
- Version numbers or tags
- 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 creating an item bank

3.1.4 Building and maintaining an item bank

An item bank serves as a centralised repository of test items that can be easily accessed, organised, and managed. It allows for systematic storage and categorisation of items by competence, level, and other relevant criteria, streamlining the test construction process. By having a well-organised item bank, testing organisations can ensure consistency in relation to formal criteria across tests, facilitate the reuse of high-quality items, and improve the assessment's overall reliability and validity. Moreover, an item bank supports ongoing development, making the process of generating new test versions that meet specific needs more efficient while maintaining a high standard of quality and fairness.

3.1.4.1 Structure and organisation

The structure and organisation of an item bank should reflect the test specifications and version development processes. Separate banks may be needed for different categories of items, for example, if there is one test version for younger learners and one for adults or if the test has separately scored sections for items testing oral comprehension and written comprehension.

Categorisation within the bank. Organise the item bank by categories that are important for the test specifications, for example CEFR level, item type, test component, or groups of CEFR descriptor scales.

Metadata and tags. Include detailed metadata for each item that reflects the kind of information that will be needed to assess item bank health (readiness for constructing test versions), or to assist with version construction, either by humans or using an automated test assembly algorithm (see 3.4). Categories of information might include learning objectives, item format (e.g. multiple choice, short answer), or difficulty level. If there are requirements to balance test versions by topic, topics or themes covered by the item should be tagged. If the test uses automated test assembly software, additional categories may be needed that will enable the software to balance items appropriately, for

example by identifying specific items that should not be administered together, or more fine-grained topic labels that will ensure variety within larger topic areas.

Version control. Implement a content control system to track revisions made to each item. Maintaining a history of changes allows reviewers to avoid repeating wordings that were found not to work, and, conversely, to revert to previous item wordings if necessary.

3.1.4.2 Technology and tools

Database software. Explore using dedicated database software designed for managing item banks. These tools offer features like filtering, searching, and reporting functionalities, making it easier to locate and analyse items. Such systems may include item development workflow management, enabling the tracking of items at different stages of development or management of item writing assignments.

Spreadsheets. Spreadsheets can be a viable option for smaller item banks. Ensure consistent formatting and utilise sorting and filtering functions for basic organisation. For certain item types (e.g. long reading passages with multiple items) it may not make sense to store the item content itself in the spreadsheet; in such cases, the spreadsheet can be used in conjunction with an organised folder structure that holds the item content.

Cloud storage. Consider storing the item bank in a secure cloud storage platform. This allows access from different devices and facilitates collaboration among item writers and test developers.

Security and access control. Implement security measures to ensure that only authorised personnel can access and modify items within the bank.

3.2 Producing materials

This section focuses on the stage of commissioning test items, which refers to the process of requesting and acquiring materials for the live test from trained item writers.

3.2.1 Assessing requirements

In order to construct a test version, 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 Sections 2.4.3 and 2.4.6 on elements that need to be balanced, as well as Section 3.4 on constructing test versions). This means that more items must be commissioned than will actually be used in the particular test version. Another reason for commissioning excess items is that some items are almost certain to be rejected at the quality control stage.

3.2.2 Guidelines for commissioning

Materials may be commissioned for a particular test administration, or to add new material to an item bank, from which test versions or pools for on-the-fly test assembly will later be constructed. In either case, sufficient time for all steps in test production must be allowed.

↘ **Considerations for commissioning:**

1. **Quantity and breakdown**. Clearly specify the total number of items needed, along with a breakdown by the categories required for the commission (specifications for one version, requirements to fill weak areas in the item bank, etc.). Be clear about how units are counted: for example, is a written text accompanied by five items based on the text to be counted as one item, or as five items plus a stimulus, or as six items? This distinction is important when setting payment expectations, and having a consistent way to count is essential to item bank management.
2. **Materials to be submitted**. As discussed in Section 3.1.2, specify exactly what must be submitted. For example, some item types will require a key or marking scheme for each

item, including the correct response(s) and any relevant marking criteria, such as specific guidelines for partially correct responses; others might require relevant information on expected performance and/or rating criteria. Also specify what audio files or visual materials must be submitted, if applicable.

3. **Schedules and deadlines**. Specify the milestones and deadlines, and if writers are contracted, state these in the contracts. The timeframe needs to accommodate the following key stages:
   - **Item writing and initial review process**. Allocate sufficient time for item writers to develop high-quality items, considering the complexity of the items and the need for revisions based on feedback (see 3.3.2, Item review).
   - **Production of audio and visual materials**. If audio materials must be recorded or edited, or visual materials produced, those will need their own production and quality control schedules, and if these are contracted to third parties, the schedules may need to be set far in advance. Ensure that the schedule for the development of the items (including dates for piloting, pretesting, or trialling) is coordinated with any such schedules.
   - **Piloting, pretesting and trialling**. Factor in time for piloting, pretesting, trialling, or a combination of these, depending on the organisation's specific aims and available resources. This phase helps refine the test items further by identifying any issues that may not have been evident during the initial review.
   - **Item bank management**. If necessary, allow time for incorporating finalised items into the main item bank or the test generation software, ensuring proper version control and security measures are in place.

3.2.3 Item submission format

Refer to the item bank structure and organisation, and provide item writers with detailed instructions on how to use tools and/or software.

1. **Item banking system**. If the item banking system allows item writers to enter items directly into the system, they will likely need detailed instructions and/or training on how to do this.
2. **Electronic files.** Provide item writers with a template that will facilitate entry into the item banking system and/or consistent review if there is no item banking system.
3. **Paper-based submissions.** If paper submissions are accepted, specify formatting preferences, such as having each item on a separate sheet of paper and the locations of the required labels (for example, item identifier, item writer's name, date).
4. **Commission acceptance form or contract**. Include a formal document for item writers to acknowledge acceptance of the commissioning process and its terms.
5. **Copyright agreement**. Consider providing an agreement outlining that the test provider will own the copyright of the developed test materials.
6. **Non-disclosure agreement (NDA**). Item writers, reviewers and vetters must sign a non-disclosure agreement (NDA) to protect the confidentiality of test content and specifications if the testing institution deems this necessary. This agreement should outline the specific information considered confidential and the limitations on its disclosure by the item writer.

3.3 Quality control

Once item writers submit their materials, a rigorous quality control process is crucial to ensure that the test items meet the standards required for the testing programme. This process is multifaceted and involves multiple steps, including, for example, expert reviews, collaborative evaluations, and practical testing. Quality control is an ongoing and dynamic process, often requiring adjustments and refinements at various stages. It is not always linear, as feedback and findings may necessitate

revisiting earlier steps. The aim is to ensure that every test item contributes to a reliable, valid, and fair assessment, upholding the integrity of the examination as a whole.

3.3.1 Quality control process

Once test materials have been submitted, they must be checked for quality. When one or more changes are made to an item or task, it should be checked again, for example, by qualified proofreaders, subject or context experts or consultants to see if it is suitable. All quality control should be documented, as it will be key evidence for the validity argument (see 1.6).

This kind of checking is essential; materials should be checked by someone other than their author. However, if the item writer is working alone, leaving extra time between producing the tasks and reviewing them, or reviewing many items at once can enhance objectivity.

The very first check should be that the materials conform to the test specifications (refer to 2.5 and 3.1) and any other requirements set out when commissioned. Feedback can also include suggestions on how the item might be changed. Constructive feedback allows item writers to review their work and develop their item writing skills.

3.3.2 Initial vetting and review process

The item review process can be conducted in various ways, typically through either an individual expert review or a collaborative expert review team; reviews may be conducted synchronously or asynchronously. This objective evaluation helps identify and address any potential issues before the items move forward in the process. Having reviewers other than the original author provides a fresh perspective, which is crucial for spotting ambiguities, biases, or areas for improvement.

Within the initial review process, it is useful to distinguish between vetting and review. Vetting can be considered as a specific stage in which test developers assess whether commissioned items meet the test specifications and decide which should be rejected outright and which can move forward. Review, by contrast, is the broader process of evaluating test materials at any point in the test cycle. In the initial review stage, vetting can therefore be seen as one form of review, focused on a pass/fail decision, whereas subsequent review activities may involve more detailed qualitative evaluation and suggestions for revision.

The evaluators are qualified test development professionals or experienced educators with expertise in the target language and the CEFR framework (or another framework against which the test is constructed). They meticulously evaluate each item or task to ensure it meets the required standards, ideally referring to the vetting and review criteria. Review comments may be documented within an item banking system, as a column of an item banking spreadsheet, or as comments on the original item document. They should include the reviewer's name (or reviewer ID, for anonymous reviews) and the date.

3.3.2.1 Vetting criteria

At the vetting stage, test developers check that each item or task fulfils the basic requirements set out in the specifications. This initial check focuses on ensuring that the item is construct-aligned, technically sound, and compliant with all essential item writing specifications before it progresses to in-depth review.

At this stage, evaluators should confirm whether the item or task:

- Corresponds to the test specifications in terms of construct, CEFR level, communicative activity/competence, and task type
- Follows the item writing specifications, including required format, text length, item type conventions, and rubric style
- Is free from immediate red flags such as obvious inaccuracy, unclear prompts, inappropriate length, construct-irrelevant difficulty (e.g. excessive background knowledge), or use of restricted content (e.g. topics on a taboo list)

- Includes required supporting materials, such as answer keys, marking schemes, prompt descriptions, or source information for input texts
- Meets any administrative or technical requirements, such as version control, metadata fields, and file-naming conventions for the item bank

3.3.2.2 Review criteria

Reviewers should distinguish between evaluating a single task or test and conducting a broader review of items within the item bank. Comments on individual tasks or items should focus on the specific features of that task or item (e.g. vocabulary, rubric clarity, key accuracy), while broader review comments should address consistency, coverage, and appropriateness across the item bank as a whole. This distinction helps maintain both item-level precision and bank-level coherence.

Reviewers should evaluate each task or item according to criteria established for the test programme. Below are examples of useful criteria:

- **Alignment with the specifications**. Does the item operationalise the construct as defined in Chapter 2? Does it accurately target the intended CEFR level, communicative activity, competence, and domain?
- **Clarity and accuracy**. Are the rubrics clear, concise, and free from ambiguity? Does the content accurately target the intended CEFR level?
- **Cognitive-level appropriateness**. Does the item elicit the mental processes, strategies, and linguistic operations expected at the targeted level (e.g. inferencing, scanning, organising discourse, producing extended language)? Is the cognitive demand appropriate and consistent with the construct?
- **Language appropriateness**. Is the language used in the item appropriate for the target test taker population and item CEFR level in terms of difficulty, vocabulary, and grammar?
- **Fairness and bias**. Is the item free from cultural bias or unfair advantages/disadvantages for any particular sub-group of test takers? Reviewers should scrutinise items for potential bias based on factors such as cultural references (avoiding obscure references or those that favour specific cultural backgrounds), gender bias (ensuring balanced representation of genders in language and scenarios), and socioeconomic bias (avoiding topics or vocabulary that might disadvantage certain social groups).
- **Taboo topics**. Were references to politics, alcohol, religion, and other topics that have been identified as problematic avoided?

Text-based items

For items referring to a written or spoken text passage, reviewers should evaluate the appropriateness of the texts in terms of:

- Length
- Topic selection
- Writing style
- Language difficulty level (this may require expert judgement and potentially referencing linguistic descriptions)
- Text types (e.g. emails, articles, dialogues) reflecting real-world language usage at the targeted CEFR level

All item types

For **all item types**, reviewers should attempt to answer the item themselves, without consulting the answer key. This helps identify issues such as unclear distractors, ambiguous wording, or multiple possible answers.

Reviewers should track:

- **Vocabulary**. Ensure there is no overreliance on overly complex or obscure vocabulary, aiming for a balance appropriate for the CEFR level.

- **Location of key information in audio files**. Ensure that the target information is distributed according to the specification (that is, not all at the beginning or all at the end of the audio file).
- **Rubric clarity**. Assess the clarity and understandability of the rubrics provided to test takers. Are the instructions specific, concise, and free from ambiguity? Do they accurately reflect the task requirements and marking and rating criteria? It might be a help for the test takers if the different test versions, including sample/practice tests, have stable and repetitive rubrics for the same type of task.
- **Key accuracy**. Review answer keys for accuracy and ensure they correspond precisely to the intended answer(s) for each item.
- **Cognitive level alignment**. Compare the item's demands with the cognitive level specified in the test specifications. Does the item require the test taker to engage in the appropriate level of thinking e.g. remembering, understanding, applying, analysing, evaluating, creating, cf. Bloom (1956)? Is the item appropriate for the level, age, and profile of the test takers?

For further guidance on review see ALTE (2005).

3.3.2.3 Feedback and revision

Following the initial item vetting and review process, offer clear and actionable feedback to item writers. In some programmes, responsibility for the item transfers to the reviewer, who does any revising. In such programmes, item writer feedback is for the purpose of improving item writer work in the next assignment.

In the event the feedback requires item writers to revise the items themselves, here are some key guidelines the reviewers can follow in providing feedback:

- **Focus on improvement**. Frame feedback as suggestions for improvement, emphasising positive outcomes for the item's quality and effectiveness.
- **Specificity is key**. Go beyond simply stating problems. Provide specific details about what needs revision and why.
- **Actionable guidance**. Offer suggestions for how item writers can address the identified issues. This might involve providing the revised item or rewording prompts, providing clearer instructions, or adjusting the difficulty level.
- **Collaborative spirit**. Maintain a collaborative and respectful environment. Encourage open communication to clarify any points and ensure item writers feel comfortable asking questions or seeking further clarification.

3.3.2.4 Review meeting and documentation

Where the review is carried out by a collaborative review team, a review meeting may be useful. A review meeting is a stage in the item development and validation process where reviewers come together to discuss the quality, suitability, and fairness of proposed test items. Item writers may or may not be present at the meeting. Ideally, all reviewers will have completed an individual review in advance, using the agreed criteria and forms, so that the meeting serves primarily to collate, compare, and reconcile individual comments. This approach ensures that the discussion is focused and evidence-based, and that differing perspectives can be constructively addressed. However, in some contexts, the review meeting may be the first opportunity for reviewers to see the items. In such cases, the organisation of the session should allow sufficient reading and reflection time before discussion begins, to guarantee that feedback remains considered and systematic. Whichever model is adopted, clarity of expectations is essential to ensure consistency, efficiency, and fairness in the review process.

- **Organising the discussion**. Use the initial review meeting notes and reviewers' comments as the focus during the group discussion. This allows for a collective review and ensures everyone is on the same page regarding feedback points. Writers may be present or not during the review meetings according to the institution's procedures. If writers are not present it is recommended that feedback be provided to them.

- **Detailed records**. Designate one team member to maintain a meticulous record of all decisions and agreed-upon changes. This record should clearly document:
  - The original item content
  - Specific revisions suggested by the reviewers
  - The final agreed-upon version of the item
  - Rationale for any significant changes
  - Any remaining questions or unresolved issues
  - Transparency for writers: ensure that item writers, even those not present during the review of their specific items, receive clear and detailed feedback on rejected or revised materials

3.3.2.5 Additional considerations

- **Handling rejected items**. For items deemed unusable, provide specific feedback on the identified flaws, stating clear reasons why the item was rejected and referring back to the specifications and specific sections in item writer guidelines.
- If writers are typically responsible for revising their own work, when an item needs substantial revision it may be useful to assign them to a more experienced writer.
- **Security and record keeping**. Be clear about version control and what needs to be retained as a record. Destroy any extra or used paper copies of materials.
- **Learning opportunity**. The feedback given to the item writers during or after the review meeting should constitute an opportunity to discuss the review as a learning experience. Keep track of recurring issues that come up in review meetings – these may inform the focus of any future training and development sessions.

3.3.3 Piloting, pretesting, and trialling

3.3.3.1 Approaches to collecting information

After the initial vetting and review of test items and tasks, it is useful to assess how well they are working with test takers or others approaching the items as a test taker would. Test providers typically use piloting, pretesting, trialling, or a combination of these (depending on their specific aims and available resources) to identify any issues that may not have been evident during the initial review. These methods may be used as part of specifications try-out (Section 2.5) or routine data collection (Section 6.1), or both.

If a sufficient number of test takers participate, this stage may also provide statistical data on the items or tasks, such as difficulty and discrimination, that enables the items or tasks to be confidently used to count toward a score.

Piloting involves asking a small number of people to complete the test items or tasks. This process can be informal, potentially involving work colleagues or other readily available individuals to complete the items and provide feedback, or formal, as in structured sessions (for example, think-aloud protocols, in which test takers narrate their thinking as they engage with each task, or post-task interviews with set questions) with people similar to eventual test takers. Their responses are analysed, and their feedback, along with any comments, may be used to further improve the items or testing system functionality (especially when using a new system). This can be particularly helpful in evaluating alignment with expected cognitive processes and competencies (see Section 2.4).

Pretesting, on the other hand, is typically conducted for tests with objectively marked items to be used in live tests and involves replicating live test conditions. Test takers who closely resemble the target population are selected, and a sufficient number of responses are gathered to allow for statistical analysis (see Appendix III for further details). This analysis provides insights into how well the items of the receptive components function, including the effectiveness of options, item difficulty, average scores, and the overall level of the test for the test taker group. Additionally, when the number of responses is large, it can reveal potential biases, measure how consistently items

assess the same construct, and provide other critical information. Items with acceptable statistics are then eligible for use in live tests to count toward a score. Responses to subjectively marked tasks, such as those testing writing and speaking, can be analysed statistically; however, a qualitative analysis based on a smaller number of responses may be more insightful. Small-scale pretesting of these productive tasks is sometimes referred to as **trialling** to distinguish it from the pretesting of objectively marked items. Trialling helps determine whether the test tasks function effectively and elicit the expected type of performance from test takers.

Unlike piloting, a pretesting or trialling programme closely resembles a live test programme, requiring similar resources and logistical arrangements. These include:

- **Sufficient numbers of test takers**. It is crucial to ensure an adequate number of test takers for reliable statistical analysis. Depending on the type of statistical analysis carried out, a minimum number of test takers or performances should be taken into consideration (see Appendix III for further details). For trialling, if the primary analyses are not statistical, the number can be small; however, it is still advisable to sample test takers from different segments of the test population.
- **Secure test distribution**. Control of paper test content or secure platforms for digital delivery, along with proctoring tools to maintain test integrity, need to be ensured (see Chapter 4).
- **Test venues** for in-person pretesting and/or robust online test functionality for remote test delivery is required (see Chapter 4).
- **Staff**, e.g. test administrators, proctors, examiners, such as clerical markers for objectively marked tasks or raters for subjectively marked tasks.

3.3.3.2 Sampling for pretesting

For the pretesting to be effective, the profile of the test takers should closely resemble that of the test takers who will take the actual live test. This approach not only simulates real test conditions but also ensures that the test takers' performance is reflective of the target population's abilities. Ideally, individual learners preparing for the live exam can be recruited for pretesting. To ensure that the learners are motivated to do their best, it is common to use embedded pretesting, in which live test versions contain a number of pretest items that do not count toward the score; test takers are not informed as to which these are. This technique ensures that the pretest population is the same as the general test population and that they treat the pretest items the same way they would treat the items that count toward a score. A drawback of this approach is that not all the items will count for scoring, so the test will need to be longer than if all items are counted. (If the test is adaptive (see Sections 2.4.3 and 4.3), the effect may not be as severe, because most of the items will be tailored to the specific ability levels.) Another drawback of embedded pretesting is that the number of items that can be pretested in this way is relatively small; for new testing programmes where large numbers of items must be pretested, pretesting will need to occur as a standalone event in most cases.

> 'Sampling of persons to help construct an instrument may be different from the sampling of persons who are to be assessed. To provide the evidence to check the operation of items, it is necessary to have persons whose responses contribute to relevant information. Thus, in the study of items in a test of proficiency, it is important to have more and less proficient persons so that the persons in this sample contribute the same information across the relevant range of the difficulties of the items. Thus, ideally, one would try to obtain persons across the whole required range of proficiency and as close as possible to being uniformly spread. Unless deliberately selected this way, samples are more likely to have a unimodal rather than a uniform distribution. If they have a unimodal distribution, such as the normal, they provide much more information about items around the mean of the persons than at the tails of the distribution.' (Andrich & Marais, 2019, p.340)

If embedded pretesting is not possible, offering feedback on performance in pretesting is a valuable way to motivate test takers to participate fully and provide responses that genuinely represent their abilities. This feedback can also be beneficial for learners and their teachers, helping them

to understand current proficiency levels and identify areas that need improvement before the live test.

However, a potential drawback of pretesting is the risk of exposing test items, which could compromise the security of future live tests. This concern can make pretesting challenging for some test providers. To mitigate this risk, the following strategies should be considered, in addition to the routine measures outlined in Section 4.6:

- **Avoid using complete test versions** in pretesting that are identical to those intended for live administration. This helps protect the integrity of the live test.
- **Build a significant time lag** between the pretesting of items and their use in live tests, reducing the likelihood that individual test takers will encounter the same items again.
- **Administer pretests with secure procedures**, especially when pretesting is carried out by test centres or schools and teachers that are contributing to the pretesting stage. Staff involved should be given clear instructions on maintaining security and may be required to sign confidentiality agreements.

It is important to note that a pretest version does not need to closely resemble the actual live test version since the focus is on pretesting the individual items, not the overall test. However, if test takers are motivated by the opportunity to experience a test format similar to the live exam — as part of their preparation — it is advisable to use a format that closely mirrors the live test.

In all cases, the pretest should be administered under conditions that replicate the live test as closely as possible. (See Section 4.3.) Ensuring that test takers are not distracted, do not cheat, and are given consistent time limits is crucial for obtaining reliable and interpretable data. Inconsistent conditions can lead to misleading results, undermining the purpose of the pretest.

3.3.3.3 Qualitative feedback

For pretesting aimed at obtaining qualitative feedback, several strategies can enhance the value of the information gathered:

- For objectively marked items, feedback from test takers and teachers is helpful. Using a structured list of questions or a questionnaire can facilitate the collection of this feedback.
- For speaking tasks involving an examiner, the examiner's feedback is crucial. The feedback from the examiner or from the rater helps determine whether the task was understood by the test takers, if it was appropriate for their level, experience, and age group, whether they received sufficient information to complete the task effectively, and if the item elicited performances at the level. For subjectively marked items and tasks, responses from test takers can reveal whether they had the opportunity to demonstrate the expected range of discourse structures, syntactic complexity, and vocabulary appropriate for their proficiency level.
- For short-answer questions, responses can indicate whether alternative answers should be accepted and incorporated into the mark scheme.
- Overall feedback on the pretesting experience can also be gathered from test takers to assess the session's effectiveness and identify areas for improvement.

See Appendix II for examples of feedback questions.

3.3.3.4 Item review after piloting, pretesting and trialling

Following a piloting or pretesting/trialling session, a review of items should be carried out with the aim of retaining, revising, or rejecting items based on the gathered evidence. The review may be carried out by a group meeting together, by a single expert reviewer, or by multiple reviewers making comments in a central location and working asynchronously.

The review process should address the following general points:

- **Determining readiness for live testing**. Identify which items are ready to be included in the live test based on their performance during pretesting or trialling.

- **Rejecting unsuitable materials**. Decide which items should be discarded due to issues such as poor performance, lack of clarity, or evidence that test takers answered based on something irrelevant to the construct.
- **Rewriting and further pretesting or trialling**. Determine which items need to be rewritten and subjected to another round of pretesting or trialling before reconsideration for inclusion in the live test.

In the review, consider the following aspects:

- **Alignment with target population**. Evaluate how well the test takers taking the pretest or trialling versions matched the characteristics of the intended live test population. This helps gauge the reliability of the evidence gathered from the analysis.
- **Engagement and accessibility**. Assess the engagement and accessibility of the tasks and topics. Review any issues related to administrative procedures during pretesting or trialling to ensure that performance on the items and tasks was not affected by any administrative or procedural problems.
- **Difficulty and alignment with CEFR level**. For single-level tests, there are generally expectations about the difficulty range that provides adequate measurement. For multiple-level tests, difficulty range expectations may be aligned to CEFR level; although there will be overlap in ranges, generally items at higher levels should be more difficult. Establish ranges appropriate for the specific test and population, and items that fall outside their expected range should be examined to see if they are truly representing the intended level.
- **Performance of individual items and tasks**. Analyse the performance of items and tasks based on the feedback and data collected from statistical analysis. For subjectively marked tasks, reviewing a selection of test taker responses is beneficial to understand how well the tasks elicit the expected performance. For objectively marked items, statistical analysis can highlight problematic items, though these findings should be confirmed and interpreted by experts. Be cautious with statistical data derived from small or unrepresentative samples; qualitative insights should be given significant weight.
- **Consistency and coherence in the item bank**. Ensure that statistical and qualitative information about items is consistently recorded and stored in the item bank. This coherent approach helps maintain the usefulness of the item bank during test construction and subsequent updates.

Best practices for the review include:

- **Detailed documentation**. Keep thorough records of any discussions, decisions made, and actions required. This documentation helps track changes and supports future reviews.
- **Expert input**. Use the expertise of item writers, markers, and/or raters as well as reviewers to provide a comprehensive evaluation of item quality. Their insights are crucial in interpreting both statistical data and qualitative feedback.
- **Follow-up actions**. Develop a clear plan for addressing items that require rewriting or additional pretesting or trialling. Assign responsibilities and set timelines to ensure that revisions are completed efficiently.
- **Regular updates**. Continuously update the item bank with new insights and revised items to ensure it remains current and relevant for future test construction.

Quality assurance in test development is a continuous and often cyclical process. Items may undergo multiple reviews during pretesting and trialling and may even require further evaluation once the test has been launched (see Sections 1.6.3 and 6.1). It is crucial to document the life cycle of each item meticulously and provide item writers with constructive feedback. This ongoing process ensures that the items remain relevant, reliable, and effective, ultimately supporting the overall integrity and quality of the test.

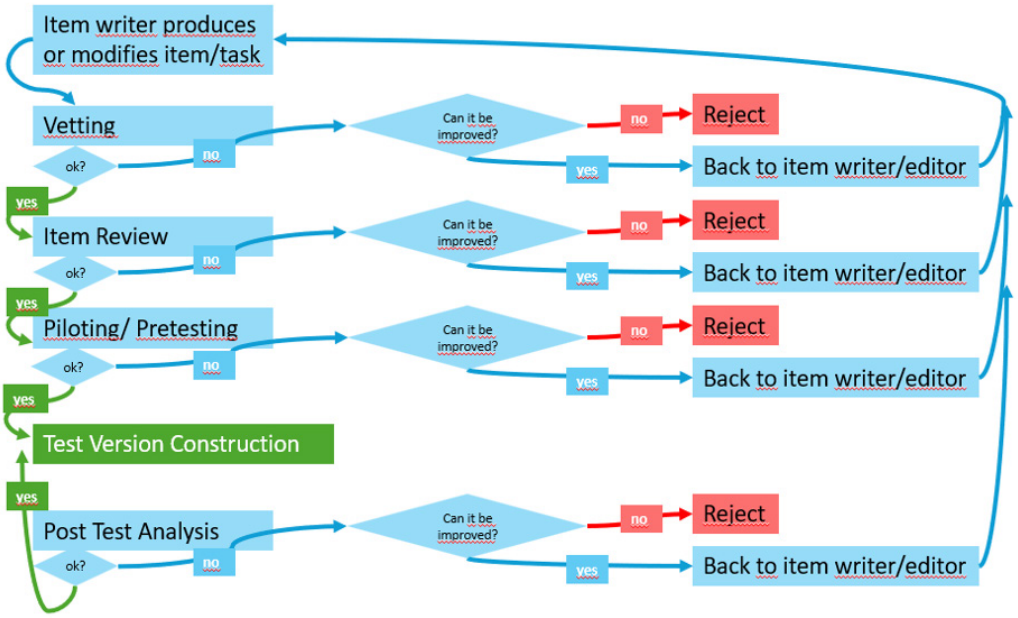


Figure 5 ***Item review cycle: Improving items with quality assurance***

3.4 Constructing test versions

This section focuses on those aspects of item development that are closely interrelated with the process of constructing test versions. This process focuses on assembling finished items into complete test versions that meet the desired quality standards and adhere to the specifications (see Section 2.4.6). Test construction may be done manually by choosing items one at a time for one specific version, or it may be done using automated test assembly software that has been programmed with test specifications. This software may be used to construct a version that is then reviewed by humans, or it may be used in linear-on-the-fly-testing (LOFT) or Computer Adaptive Testing (CAT), in which the algorithm is tested through multiple simulations evaluated by humans, but ultimately selects items and tasks for versions without human intervention. Such algorithms typically include the test specifications, but also include requirements that mimic human judgement, for example including lists of items that cannot be administered together (see Chapter 4). In either case, some basic principles apply to considering details beyond the specifications, discussed below.

3.4.1 Balancing content and variety for CEFR alignment

A crucial aspect of test construction is ensuring a balanced and varied distribution of relevant content domains and an appropriate distribution of linguistic features relevant to what is being tested. Some of this distribution may be included in the specifications, for example required numbers of items in particular content domains or at particular CEFR levels (for multi-level tests). At a more detailed level, however, ensure an appropriate distribution of linguistic features relevant to the CEFR level(s) being tested, so that the items test different aspects of the levels within the constraints of the specifications. (Refer to familiarisation, specification and possibly standardisation phases of CEFR alignment: see the Council of Europe (2009) Manual and the British Council et al. (2022) handbook).

To facilitate this balance of detailed characteristics, consider tracking the following for each item:

- **CEFR level descriptors**. Identify the targeted CEFR level for each item and link it to appropriate descriptor(s) for the communicative activity.
- **Topics**. It is helpful to categorise items based on the topics they address. This enables the monitoring of topic distribution and ensures that a variety of subjects are covered within each test version at a specific CEFR level.

- **Lexico-grammatical demands**. If the test includes items that target specific linguistic forms, such as those in which knowledge of a particular word or word form is essential for answering correctly, label the items with these forms to avoid testing the same one twice in one test version.
- **Text types (for text-based items)**. For item types referring to written or spoken texts, it may be useful to categorise items based on the type of text (e.g. emails, articles, dialogues) they use or on the CEFR descriptor scale(s) they refer to (e.g. reading correspondence, instructions, listening as a member of a live audience, etc.).
- **Task focus**. Identify which linguistic task(s) the item is intended to operationalise — for example, whether it targets understanding the main idea, following simple instructions, extracting key points, or interpreting a speaker's attitude.
- **Audio features**. For oral comprehension components, it may be useful to identify features that may affect performance, such as speaker gender, speech rate, accent (if applicable), amount of background noise, or register (formal vs. informal language) and categorise items accordingly to ensure a balance of these elements across different test versions.

By systematically tracking these elements, it is possible to construct test versions that offer a variety of content within the categories required by the test specifications.

3.4.2 Other test features

Item overlap. Ensure that no items provide clues to other items (for example, a linguistic form tested in one item being used prominently in another). Note that this is not quite the same as two items targeting the same linguistic form: here, one item may target the form, while the other uses it but may not focus on it. For example, there may be an item targeting a particular collocation, and another item based on a text that includes that collocation; here, even if the collocation in the text is not tested, it provides an example that will clue the correct answer to the item in which it is the target. This type of overlap is especially important to capture when using automated test assembly software, in which case it may be necessary to create or adjust tags for specific items that cannot be administered together.

Level of difficulty. Use the statistical information stored in the item bank after pretesting to ensure the mean item difficulty for the version is within a specified range, so that different test versions and components are comparable. Sometimes the mean item difficulty range is part of the specifications, but if it is not, it is still useful to ensure that there is not too much variation. Depending on how the version's components are put together, it may be useful to specify difficulty ranges for different item types, each component, or the version as a whole. If IRT is used (see Appendix III), it may be more useful to use a Test Information Function rather than mean difficulty. Note, however, that pretesting is normally carried out with small test taker numbers, e.g. 150 to 200, so that difficulty values may have large confidence intervals and should later be checked by post-test analyses (see Appendix III for further details).

3.4.3 Quality control for test versions

Section 3.3 on quality control concerned mainly the individual items and tasks. When putting together a test version, additional checks on items and tasks may be conducted, but it is also important to check the version as a whole, particularly when there are specifications for collections of items. For example, individual input texts may have word count or time limits, but there may also be specifications for word count for the test component as a whole. For components with audio files, there may be specifications for the length of pauses between one audio file and the next. Ensure that there is a quality control stage to check for these aspects of the test specifications.

3.5 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 the initial item review?
- Will review be done by individual reviewers writing comments, or in a group meeting?
- How will feedback to writers be addressed?
- How will it be possible to pilot, 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 version construction purposes, for test writer training, etc.)
- What will be the role of statistical analysis? (e.g. in establishing the mean difficulty and difficulty range of the test version, or of items at particular CEFR levels)
- When making decisions, how important will statistical analysis be in relation to other information?
- Who will be responsible for the item review cycle until the item is retained?
- Who will be involved in the activity of test version construction?
- In what way will automated test version construction be monitored?
- Which variables need to be considered and balanced against each other? (e.g. level of difficulty, topical content, range of item types, etc.)
- How will integrated tasks be specified and scored to ensure that the intended competences are the focus of assessment?
- Will the constructed test version be reviewed independently?
- How will the constructed test version be matched to other versions of the same test or made to fit within a larger series of tests?

3.6 Further reading

For item writer guidelines, see ALTE (2005).

For the analysis of test tasks, see ALTE Content Analysis Checklists and Checklists for Single tasks (https://www.alte.org/Materials).

Linguistic descriptions of some languages that relate to the CEFR are available: Reference Level Descriptors (RLDs) (Beacco & Porquier, 2007, 2008; Beacco et al., 2004; Glaboniat et al., 2005; Instituto Cervantes, 2007; Spinelli & Parizzi, 2010; https://englishprofile.org/).

List of RLDs assessed or deemed to comply with the Production Guide:

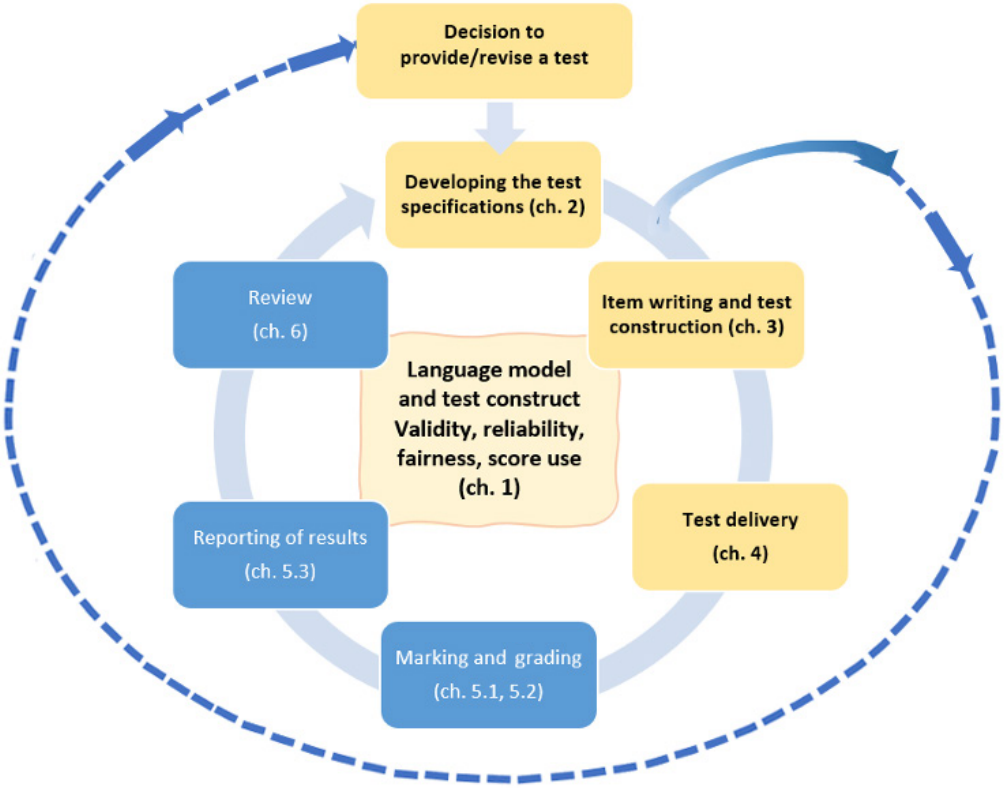
https://www.coe.int/en/web/common-european-framework-reference-languages/reference-level-descriptions-rlds-developed-so-far

Next steps

Chapter 3 built on the foundations established in Chapters 1 and 2 by outlining how test specifications are operationalised into concrete test content. It provided a structured, practical framework for producing, managing, reviewing, and assembling test materials.

Chapter 4 will take the test materials developed and quality-assured in Chapter 3 and move into the operational phase of the Testing Cycle by outlining how tests are delivered.

4 Delivering tests



4.1 Introduction to test delivery

Test delivery is the stage in the testing process in which the test takers must be given every opportunity to prove their proficiency. In order to demonstrate their knowledge and skills, test takers are expected to complete assignments, speak, write, listen, summarise, rephrase, etc. Each of these assignments involves the collection of data that allows for an assessment of their competence. Thus, the main aim of the test delivery process is to collect accurate and reliable information about the ability of each test taker, regardless of the exam medium (paper or digital). (See Sections 1.5.1 and 1.5.2.)

Test delivery has evolved from a simple, paper-and-pencil-based process to a more complex, multi-layered and largely automated process in which computers or other digital devices and software or testing platforms play an increasingly prominent role.

4.2 Test taker data collection

Before going into the process of delivering the real test, it is useful to discuss the collection of important data about the test taker. The collection of this information might happen during registration, but it also might happen in a post-exam questionnaire, or in some cases by answering questions in the test. Not all data about the test taker is needed to administer the test; we need information about potential accommodation for special needs ahead of time, but information needed for DIF analysis, for example, need not be collected in advance.

4.2.1 Data collection on test taker background

In order to offer each test taker fair and equal opportunities to demonstrate their abilities, the necessary (and *only* the necessary) personal information about the test takers must be obtained. With that information, a suitable environment can be created or adjusted to facilitate equal opportunities for all test takers. Each testing programme should determine which information is necessary, as this may vary depending on test purpose, context, and level.

Test taker background information

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 enroll, information about the test takers needs to be collected (e.g. age, gender, linguistic background, relevant prior knowledge or educational context, target language use domain). Information regarding email, phone number and home address might be necessary.

The data which is collected is relevant for the following steps of testing:

- For the **delivery** itself:
  - For face-to-face administration: the number of participants in the written and oral exam is important – examination rooms, exam papers, logistical arrangements, invigilating personnel, examiners and possibly recording instruments for the oral exam need to be arranged
  - Information on possible test takers with special needs will determine necessary arrangements
  - Digital administration in testing centres requires similar arrangements; test administrators need to be sure that they have the number of computers, tablets or other devices the test takers will use; for remote administration the platform, internet connection and remote invigilation need to be prepared and verified
  - If there are invigilators supervising the testing process online, their number needs to be coordinated with that of the test takers in terms of possibility of coverage
- For the phase of **marking and grading**: it is important to know in advance how many markers and raters are necessary and for which levels, etc.
- For the phases of **validation and post-exam analysis**: the information will contribute to the evidence for the validity of a test (e.g. was the test taking population of the same profile with the intended one?); it will also strongly determine the granularity with which psychometric and other analyses can be conducted: the more data is available about the test takers, the more refined the analyses and conclusions will be (see Chapters 5 and 6)
- For **communication** of results: test providers and/or administrators need to know in advance the contact details of the test takers in order to inform them about their results; the test takers will use email addresses and log-in details in order to access the information provided online, including possibly the results; home address will be necessary in case paper certificates are sent to the test takers by the testing centres

Test providers need to keep in mind that the European General Data Protection Regulation (GDPR) invites us to be parsimonious in personal data collection. It is good practice to just collect data which is relevant and effectively used for test administration and analysis, and test takers should be allowed not to answer questions that are not critically necessary for the services that are offered. 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. (See also Section 1.5.1.5.)

Test takers need to be informed that the data which is collected may be used to improve the exam and the whole testing system. Consequently, they will be used for the benefit of current and future test takers.

4.2.2 Data collection on test taker ability

The main aim of the test delivery process is to collect accurate and reliable information about the ability of each test taker, regardless of the exam medium (paper or digital). Given that the test content has been properly assembled (see Chapter 3), the test delivery process must ensure that factors that interfere with the measurement of ability as envisioned by the construct (see Section 2.4.1) are minimised.

The big challenges in test delivery are thus not content-related, but logistical or linked to test takers' exam experience (e.g. digital skills) and user-friendliness of the exam delivery devices.

Test providers must ensure that:

- All test materials (in physical or digital form) are available in the right place at the right time, digital platforms are up-to-date, access codes are available, etc; if the test has a speaking part, the administration needs to be prepared in terms of: enough rooms for avoiding lengthy waiting, examiners ready to administer the questions (if human examiners are used), recording and storing instruments, if necessary
- A test taker's performance on the test is determined by their language ability and influenced as little as possible by irrelevant factors, such as a noisy environment, or technical abilities (or problems)
- Delivery conditions are adapted for test takers with special needs
- The delivery process is consistent, fair and fraud-free, creating equal chances for all participants; fraud not only influences the individual results of one test taker, but also might have an impact on the total test taking population due to the unfair advantage and the distortion of the statistical and psychometric values
- The test takers' responses are collected efficiently, securely and are available for the next stages of marking and grading

These tasks are important whether the test is on a large or small scale. Something as simple as checking the suitability of a room or the correct functioning of the test devices before the test can make an important difference. Where test providers are not directly involved in administering the test, the test centres must be supplied with detailed regulations on the necessary administrative steps.

Language tests often include an oral examination component (production and/or interaction). While more preparation might be necessary for the oral test taken in person, requiring more controlled delivery conditions, and arrangements made in good time, the increase of test delivery means and tools (e.g. for taking oral tests remotely) facilitates the administration of the speaking components of exams.

This variety of options is beneficial, creating more opportunities for test takers to participate, provided adequate measures are set up to guarantee fairness.

Besides their obvious use for scoring, data collected in this step can be used for research purposes or for rater and examiner training (see Sections 6.1 and 5.2). The latter concerns samples of test takers' written work and audio or video recordings of the oral part of the examination. A way for regulating the use of such data, in terms of duration of storage and pseudonymisation/anonymisation, information for the test takers, obtaining their consent, and the consequences of consent not being given, needs to be defined. Taking part in the test should not depend on consent to use data in such a way.

4.3 The process of delivering tests

4.3.1 Before test delivery: test taker registration and informing the test taker

Registration

Test taker registration is an essential operation in the process of test delivery, with multiple purposes. First of all, it is the action which gives the test taker access to the testing process. It is

also the opportunity to provide the test taker with information about the testing process, and, as mentioned before, to collect data. All these are detailed below.

Whether it is taken on paper, face to face or in a digital form, the actual participation of the test taker in the examination process starts with registration. This can be done online or in person, centralised by the exam provider or through testing centres. Registration may be carried out by independent bodies, such as a Ministry of Education. The test provider should make every effort to ensure that access to registration is uniform for all test takers.

Inform test takers

In addition to collecting data, registration is also a good opportunity to provide information to test takers, possibly in addition to what is already available on the websites or in other informative materials. They might have extra questions about the conditions of registration, legal terms and conditions, rules when sitting the exam (e.g. to acknowledge items that may not be brought into the test centre, or to agree not to engage in cheating), the possible arrangements for test takers with special needs (see Section 4.3.2.4), the possibilities for appeal, how and when to request special assistance, the time and modality of communicating the results, 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 website or through a standard email, as well as through permanently available support systems.

4.3.2 The process of delivering tests

The test delivery process is represented in Figure 6. Depending on whether the test is paper-based or digital, the scenario splits into two at some steps in the process.

The paper workflow requires logistical processing that can be done online for most digital tests. Nevertheless, the various steps run remarkably parallel.

The process of test delivery for language assessments, whether paper-based or digital, follows a sequential workflow that begins with preparing empty test forms and then arranging the time and place for test administration. Once logistics are established, the necessary materials are prepared — either physical paper-based materials are sent to test locations, or computers, tablets or other devices (with online access) are made ready. For paper-based exams, the exam version can also be made available to the testing centres through electronic means (e.g. drive, platform) from where the testing centre personnel will download and print it in the necessary number of copies. On test day, test takers are identified to ensure proper authentication, after which they take either the paper test or digital test depending on the format. Following administration, the integrity of materials is checked, and paper materials are returned or digital test data is verified for security, culminating in the final stage where test data (from completed tests) is ready for scoring. This parallel pathway approach accommodates both traditional paper-based testing and modern digital testing while maintaining consistent quality control and security measures throughout the delivery process.

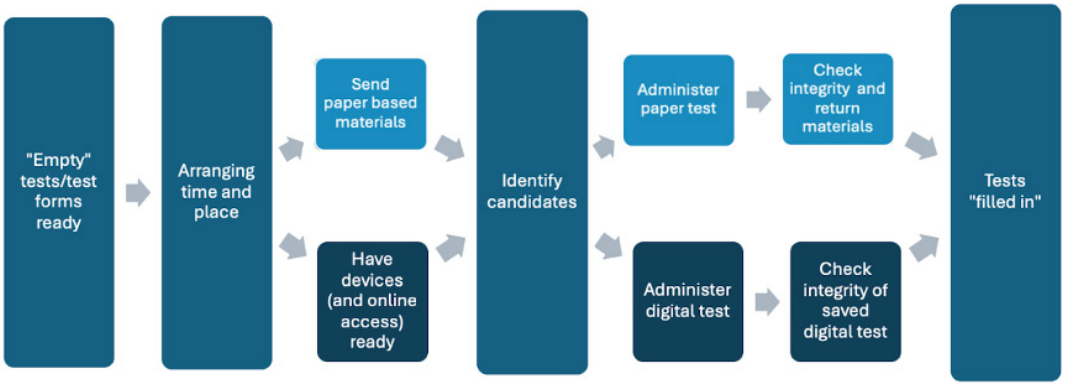


Figure 6 ***The test delivery process, splitting for paper-based or digital testing***

After this delivery phase, marking and grading can begin, although it must be said that many tests perform marking and grading at least partially instantly and automatically (see Chapter 5).

The next section discusses the elements that are common to both paper-based and digital test delivery.

4.3.2.1 Arranging time and place

Test delivery today is a multimodal process: language tests are administered both on paper and digital, but in addition to format, time and place play an important role in determining a specific type of test and the associated guidelines.

Test takers can take a test:

- In fixed/permanent or provisional/temporary exam centres (test delivery can also happen remotely, from the workplace or home)
- At pre-arranged fixed times or as 'walk-ins', which can be taken at any time in an online setting

Each of these options comes with specific and appropriate supervision settings, to guarantee fair and fraud-free testing.

Language testing delivery models can be organised along two key dimensions, time and place, creating four distinct testing scenarios (Figure 7). Traditional test centres operate with fixed times and fixed locations, requiring test takers to be on-site and pre-registered for scheduled administrations. Scheduled remote testing maintains fixed time requirements but allows test takers to test from home or work through online proctored systems with live supervision or automated proctoring. On-site open/free testing offers flexibility in timing while keeping the fixed location requirement, enabling walk-in test centres and testing windows at physical locations. Finally, flexible remote testing provides maximum convenience by allowing test takers to choose both when and where they test, utilising online proctored, AI-proctored, or unproctored formats accessible from home or work environments. The framework shows how language testing has evolved to balance security, standardisation, and accessibility through varying combinations of temporal and spatial constraints.

Test centres

Test delivery is usually done based on a schedule, in which certain venues are made available for test deliveries during a certain period (short or long).

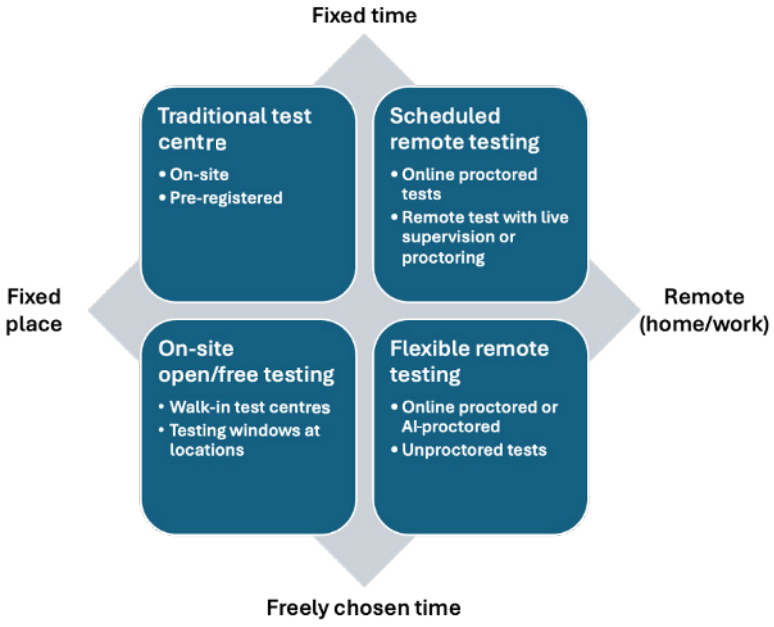


Figure 7 ***Test delivery options based on fixed or free time and place***

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 the school or the testing centre 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
- Capacity of ensuring the staff necessary for different roles (e.g. invigilation, management, administration, examining in the oral part)
- 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.

The features to be checked include: ambient noise; internal acoustics (especially for tasks requiring listening and/or speaking); 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; a waiting area for test takers; secure storage facilities for exam materials before and after the administration; secure storage facilities for test takers' personal belongings (notably smartphones).

Administration of oral exams needs to be carefully considered, since this might take significantly more time than the written components, in case there are large numbers of test takers. Enough administration rooms, examiners and recording instruments (if necessary) need to be available for administering the oral part in good conditions.

Venues which are found to be unfit for purpose or organisations which make serious errors in administration will be warned to ensure the agreed conditions and eventually may be removed from the list of possible venues or collaborators.

With digital tests, it is not enough to simply inspect the rooms; the readiness of the digital infrastructure must also be tested: computers (or other devices used to take the test), screens, peripherals (with special attention to headphones), bandwidth, etc. It is crucial to perform a full and representative test on each digital device at regular intervals. This is the only way to sufficiently guarantee readiness.

Clear guidelines should exist to ensure appropriate test centre accommodation, so that test takers can focus on taking the test and demonstrating their knowledge and skills rather than on other matters.

Remote online test delivery (with proctoring)

Where sufficient connectivity is available, tests can also be administered remotely, with the test takers sitting in a location of their choice and taking the test via an online platform.
In the case of high-stakes language tests, remote testing will almost always involve the use of remote online invigilation, also called (remote) proctoring.

However, possible problems with remote examinations need to be understood and addressed. Some organisations that adopted remote examination during the COVID-19 pandemic returned to in-person examination once the confinement was over. They needed remote administration to be further researched in order to understand implications like the risk of fraud.

Remote examining involves the collection of a large amount of personal data, including video and audio recordings. It is advised that only essential information be collected to comply with strict privacy standards, reflecting a commitment to safeguarding personal data in line with legal requirements, such as GDPR and the EU AI Act. A policy on deleting data that is no longer needed for the purpose of test administration must be defined. The purposes for which data, including recordings, is stored must be clearly outlined.

4.3.2.2 Identifying and authenticating test takers

Special mention must be made of the identification of the test taker. In many test settings, the test taker and the invigilator or the examiner who administers the (oral part of the) exam do not know each other and it is necessary to have a procedure in place whereby the test taker can be correctly identified as the person authorised to take the exam, usually with some form of official proof of identity. Both with paper-based delivery and digital tests, both on-site and remotely, there is the risk of a test taker impersonating someone else. There is also the risk of making a mistake in the delivery of individualised exams, which would have an impact on the results of the test takers affected.

In the context of language exam delivery, identification and authentication serve distinct but related **security functions**.

- **Identification** is the process of establishing who the test taker claims to be; it answers the question 'Who are you?' This typically involves the test taker presenting credentials such as a government-issued ID, passport, or registration confirmation, and having their name verified against the registration list to confirm they are the person scheduled to take the exam.
- **Authentication**, on the other hand, is the process of verifying that the test taker actually is the person they claim to be; it answers the question 'Are you really who you say you are?' This involves matching the person physically present (or online) with the identification documents provided. In essence, identification establishes the claimed identity, while authentication confirms the truthfulness of that claim, working together to prevent impersonation and ensure that the correct individual receives credit for the test performance.

Especially with large numbers of test takers, or when the oral examination is taken separately from the written one, identification of the test takers will have to be repeated. If the oral exam is video- or audio-recorded for asynchronous double rating or for responding to appeals, it is useful for the test takers to introduce themselves in the recording for easier and more secure identification by the raters. In these conditions, later checks can be performed, for example in comparison of the written and oral production of the same test taker in case a significant discrepancy between the results is registered.

With online proctored tests, it is also common for the test taker to show proof of identity, followed by a photo taken via webcam, which allows both to be compared for correct identification. For privacy reasons, caution should be exercised when conducting this ID check. If this check is carried out by a third-party platform, a signed agreement should clearly limit the storage and use of this data.

As person identification and authentication are both highly personal data, all GDPR guidelines should be strictly followed for data treatment and storage.

4.3.2.3 Administering the test

Guidelines for testing in test centres (for both paper-based and digital tests)

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. This can be ensured through training and/or documentation. 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 event of a potential danger to test takers and personnel, their safety takes priority. While the security of the examination materials must be ensured as far as possible, the protection of the test takers and personnel must be prioritised. 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 telephone 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 details 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. Procedures should be clear for as many foreseeable situations as possible, and test takers should be informed as soon as possible of what will happen in each situation, or of the timeframe within which a decision on the solution to that case will be made and communicated.

4.3.2.4 Accommodations for test takers with special needs

Requests for special arrangements should be properly assessed and, where appropriate, some form of assistance (e.g. read-aloud software) and/or compensation (e.g. extra time) 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 how long before the test the request should be received.

Different types of disabilities or needs will have to be considered, each of them requiring specific measures.

Physical disabilities

Test takers with visual or hearing problems should be provided access to the testing materials in a way that mirrors the way in which they could access real-world materials. This may include enlarged print, material in braille script, special test materials to replace any visual input, using a braille writer in the writing exam, screen readers, or listening to input in higher volume (earphones may be provided), lip reading in an oral examination, and extra time as needed. In all cases the test provider should assess how these compensatory policies affect the construct, since the construct typically is specific for the channel (oral or written). If test takers make use of their own electronic devices, it must be ensured that they cannot access the internet, electronic dictionaries or the like. If no compensatory policy can be found, test takers may be exempted from parts of the test. It should be assessed in what way any such measures need to or can be mentioned in the certification, taking into account the test takers' entitlement to equal rights and data protection.

For test takers who are unable to walk and need help to take their place in the examination room, assistance needs to be provided. For test takers who use wheelchairs or other mobility devices, care must be taken to ensure that the devices are not concealing forbidden materials, but the dignity of the test taker must be maintained. Test takers with physical disabilities affecting their ability to enter responses on the test may need alternative means, such as voice to text technology, keyboard rather than mouse navigation, or an assistant to enter responses for them. Again, it is crucial to assess how any such means affect the construct; for example, if a test requires written input, such as a sentence or an essay, someone who is speaking the response may not be interacting with the language in the same way as someone who is writing. If correct spelling is an element of the marking or rating criteria, voice to text technology may be inappropriate, for example.

Learning disabilities

A common example of accommodation for learning disabilities is the allocation of extra time, both on paper and digitally. It is a simple but often effective measure, but should be granted with care to those who need it, and only to them, so as not to create inequality. In the case of dyslexia, modified criteria for the assessment of writing may be issued.

Temporary disabilities

Test takers might have temporary health issues like a broken arm or leg. In case taking the examination is not absolutely urgent for them, this can be postponed. However, if taking the examination cannot be postponed, arrangements should be made by the testing centre. In many cases, the same kinds of accommodations can be used as for those who have permanent disabilities. For example, if a test taker has a broken arm, an assistant might be appointed to help with writing.

The test providers, directly or through the mediation of the testing centres, will respond to the special needs of test takers to the extent of their possibilities. Differences in administration must not, however, advantage some test takers in comparison to others.

4.3.2.5 The specific case of oral tests

Many language tests include an oral examination, which implies different resources and measures to be taken. The rooms need to be prepared and sufficient examination spaces and adequate scheduling need to be available to avoid long waiting times for test takers. The input materials need to be prepared in advance according to the number of test takers. For example, sufficient input materials and prompts need to be available so that if the content is transmitted to test takers who have not taken the oral test yet, the knowledge will not unfairly advantage them.

There are schematics in which the flow of test takers is organised in such a way that test takers who have already taken the test do not come into contact with participants who are still waiting for their turn. Arrangements of this kind can have a major impact on the need for item variation.

If human examiners are used, they need to be prepared and trained for administration (see Chapters 1 and 2 and Appendix I). They need to have guidelines for the interaction with the test takers and to be instructed how to react in different situations (e.g. how much they are allowed to intervene when helping the test takers if they have problems understanding the task or producing their response). If rating is done live, the rating instruments need to be prepared. If the production is recorded on audio or video, the recording instruments need to be prepared and checked in advance.

In case the examination is administered in a digital form, examiners and test takers need to know in advance the administration procedures (e.g. test taker identification, indications on recording, use of the platform, how to react in case connection is interrupted, etc.).

4.4 Paper-based test delivery

This subchapter focuses on some essential points of paper-based tests, the upper flow in the overall scheme.

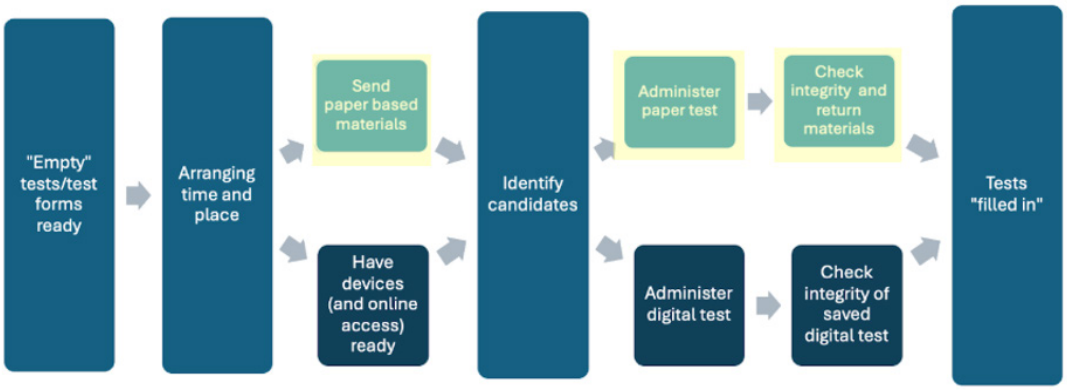


Figure 8 ***Paper delivery***

4.4.1 Sending test materials

If not created or printed in the test centre itself, paper materials 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. For security reasons, some test providers prefer to send the materials the day before the exam. These decisions will depend a lot on the local context and will be taken from case to case, with availability, but also security criteria in mind.

Administrators should check the content of the dispatch 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.

Given that many exam locations have printing facilities, electronic versions of test materials are often transmitted to the exam locations, where they are printed and copied on site. Secure and strict procedures for receiving and handling the documents are also recommended in this case. All documents involved should at least be protected with a regularly changing password.

Only clearly defined and well-instructed personnel should be allowed to handle the materials while preparing them for the examination session. The papers must not be accessible to anyone else before they are used by the test takers.

4.4.2 Administering paper tests

Basically, the administration of paper tests is simple and only requires sufficient and well-trained invigilators, in proportion to the number of test takers.

Refer to Section 4.6 on fraud for appropriate measures.

4.4.3 Returning materials

Paper-based tests must be collected correctly, in accordance with the necessary security procedures. Care must be taken to ensure that the sometimes busy or chaotic collection process does not provide opportunities for fraud. It is also important to check the integrity of the collected copies:

- Has everything been submitted?
- Have test takers filled in all necessary data (e.g. their names or codes on the answer sheets or other required test taker information)?
- Have all copies been identified or are they identifiable?
- Have all draft sheets been submitted or destroyed?

Examination papers should be packed and returned to the testing body. Any relevant documentation, such as attendance registers and room plans need to be sent to the testing body, according to established procedures, usually 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. The materials can also be scanned and sent to the test provider through a secure system.

Test providers need to define a way of dealing with any materials left over after a test, such as unused test papers, or used or unused draft sheets. Test centres should be provided with these regulations, and a way to check whether they were fulfilled must be set up.

4.5 Digital test delivery

This section focuses on some essential points of digital tests, the lower flow in the overall scheme.

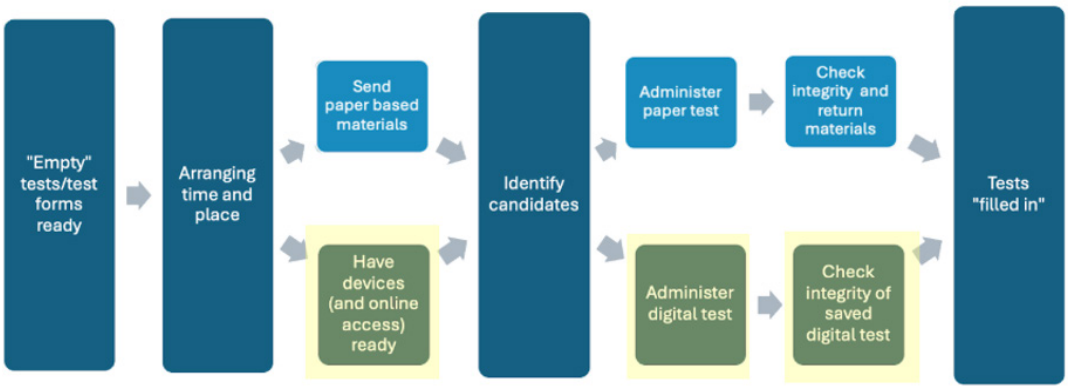


Figure 9 ***Digital delivery***

4.5.1 Digital test construction

Since the rise of digital testing, new possibilities have arisen that cannot be realised with paper-based tests. The choice between paper and/or digital must therefore be made at the earliest possible stage (see Chapter 2), because it largely determines the opportunities and limitations and has a major impact on the item writing processes (see Chapter 3) and on the test composition and the processing of results.

It is worth noting that both paper-based and digital delivery options may be available for the same test. Test organisers need to consider the implications of this and what needs to be done to ensure equity for all test takers, regardless of the delivery method.

It is impossible to summarise and discuss all the possibilities for digital delivery here, but we will explain the most prominent ones: the possibilities for test construction.

The points that are made in Section 3.4 about the construction of a test version apply to paper-based as well as digital versions. Digital delivery, however, allows a variety of modes of presentation which in paper-based testing is only available in a very restricted way. In a paper-based test the number and order of the items and tasks is generally fixed. There are ways to introduce more flexibility, e.g. to prevent cheating it is possible to provide two variants of a version with a differing order of items, provided that equality of the variants is ensured and that which test taker had which variant can be securely traced.

Digital delivery allows for the full spectrum of technical possibilities:

- The test version can be the same for all test takers. This is known as **fixed item selection**. Each test taker taking the same exam will receive the same tasks and/or items, which can be played:
  - in the same order (the method mostly used in paper-based testing)
  - or in a random order for each test taker (the method most helpful when test takers are in the same room)

Another possibility for this test construction is the one in which the choice of items is fixed, but in which different versions are created. In this case, the item variation is not 'in' the test itself, but the variation is offered by presenting different test takers with different variants (of tests with a fixed item selection).

This variation may also occur with paper-based tests, though it may be more challenging for the listening component.

- Test construction can be done **on the fly**. In (linear)-on-the-fly item selection, a test template is compiled with a framework in which a number of items are 'pulled' from a pool of available validated items and presented to the test taker based on various metadata (topic, difficulty, use of certain media types, etc.). Each test taker taking an on-the-fly test receives a different, random set of questions, but each time based on the same underlying pattern. It goes without saying that this is only possible with a sufficiently calibrated item set with the necessary metadata. On-the-fly testing is impossible on paper.
- The composition of tests can also be **adaptive**. In adaptive tests, the selection of the next item or component is based on the score obtained in the previous item(s) or test component(s). There are several other forms of test adaptivity:
  - **Branching adaptivity**. Here the item selection is determined by a test taker's 'choice' and not by score. For example, a test taker is offered three texts and can choose which of the three texts they want to answer questions about.
  - **Score-based adaptivity**. Here the test taker's score determines the further course (or the further item choices) of the test. An adaptive stage can be made item by item or chapter by chapter. The most common form of adaptivity is where the level of difficulty of items is adjusted based on the results of the previous (set of) items.
  - **IRT-based adaptivity**. Item Response Theory (IRT) places the test takers and the items on a continuum, with the position determining the probability of a test taker correctly answering an item (See Appendix III). In an IRT-adaptive test, an algorithm is used to present items that are

close to the test taker's ability. This is a very dynamic process in which new calculations are made after each item to guide the subsequent selection of items.

None of these adaptive deliveries are readily available on paper.

Finally, it is worth noting the possibility offered by digital tests to show test takers **immediate feedback and results**. This can be done after each item or immediately after the end of a test, at least with some exam components, especially those that are objectively marked (see Chapter 5). This is impossible on paper and opens up completely new perspectives for motivation, the testing effect on learning (Dirkx et al., 2014) and the (immediate) scheduling of retakes or follow-up tests.

4.5.2 Digital test delivery

Whereas in paper-based tests the test taker has a great deal of freedom in completing the test, a digital test can be much more tightly guided or directed: digital tests can be set up in different ways. Time settings, navigation settings (completely free, linear, chapter by chapter, etc.), with or without help messages, with or without on-screen tools (text-to-speech support, online dictionary, spell checker, etc.) – the possibilities and combinations are endless.

The user-friendliness and proportionality of these types of instruments are of great importance: the question must be asked whether or not the provision of certain settings or tools offers a sufficiently justifiable contribution to the testing.

Three important rules of thumb are crucial:

1. A reliable testing platform should do as much **tracking and logging** of the actions carried out by the system and by the test taker as possible (for example, record every interaction and the exact timestamp) so that a history of a test-taking session is available, which can be used to demonstrate step by step how the test progressed. This traceability is without a doubt one of the absolute added values of digital testing. One important use is research, e.g. into test takers' strategies; another is the monitoring of cheating, e.g. by comparing answer times for each item.
2. Test takers should be given the opportunity to familiarise themselves with the testing platform at least once before they start the real test. This can be online beforehand or in a practice test that does not count towards the final score.
3. Each type of (technical) interaction or question should have been practised at least once in a non-binding context, so that the test takers can familiarise themselves with the environment without stress. This allows them to demonstrate their abilities to the maximum in the actual test.

4.5.3 Digital test data integrity

Digital tests do not normally involve collecting and shipping paper documents. The data is automatically stored on the digital devices or online servers and often automatically delivered as well. Nevertheless, it is also advisable to build in a check of the data integrity of an exam session: a reliable test platform must make it clear to the test supervisor that the data needed to assess a test taker has been correctly stored for further processing in the system. Only then can a digital exam be terminated and the test taker sent home or be completely dismissed. If technical problems have potentially compromised the integrity of the exam data, the test provider must be prepared to implement solutions. They also need to inform the test taker about what happened and explain what the solution is in a situation like this.

4.5.4 The specific case of remote proctored test delivery

Online proctoring, or the remote supervision of exams taken outside traditional settings, has become a key tool in modern testing. During a remotely supervised exam, multiple monitoring methods can be employed simultaneously, such as recording the test taker's screen, capturing video via a webcam, and using a smartphone camera for a comprehensive 360° view of the surroundings. These features help detect potential irregularities, ensuring integrity even at a

distance. Exam sessions are recorded and may be overseen in real time or reviewed later, either through automated (AI-) systems or by trained personnel. Such platforms typically operate around the clock, with core functionalities accessible 24/7 and additional services, like live human monitoring, available by arrangement.

It is important to keep in mind that remote proctoring offers a range of new possibilities, but at the same time creates an additional, technical barrier for many. Remote testing and proctoring can therefore be a blessing for some, offering a solution to problems like transport and time, but for others it can be an insurmountable obstacle to taking a language test.

A fair principle is that test takers should be able to choose between more than one method of test delivery and not be limited to only remote proctored tests.

4.6 Fraud prevention and detection

Safeguarding integrity represents one of the most critical challenges in modern language test delivery, requiring comprehensive measures across both paper-based and digital formats. Test providers recognise that fraud not only compromises individual results but also undermines the validity of the entire testing system, creating unfair advantages and distorting statistical benchmarks that affect all stakeholders.

4.6.1 Paper-based test fraud prevention

Physical security measures form the foundation of fraud prevention in paper-based testing environments. Test materials must be stored in locked, secure facilities with restricted access limited to authorised personnel only. Chain-of-custody documentation should track every movement of materials from printing through delivery, administration, and return. Tamper-evident packaging and serialised test booklets help identify any unauthorised access, while strict protocols governing who can handle materials and under what circumstances minimise opportunities for pre-exposure or theft.

Room arrangements and invigilation provide the primary line of defense during test administration. Examination rooms should be configured to maximise visibility, with adequate spacing between desks — ideally at least one metre — to prevent visual copying. Invigilators must be trained to recognise common fraud indicators, including suspicious body language, excessive glancing at neighbours, unauthorised materials concealed in clothing or on desk surfaces, and the use of sophisticated devices such as miniature cameras, hidden earpieces, or smartwatches capable of storing information or receiving external communications. The ratio of invigilators to test takers should be sufficient to maintain constant vigilance, typically no more than 30–40 test takers per invigilator.

Procedures before starting the test establish clear expectations and reduce opportunities for cheating. Test takers should be required to store all personal belongings — particularly mobile phones, smartwatches, and electronic devices — in designated secure areas outside the examination room. Transparent bags for permitted items like water bottles or tissues make unauthorised materials immediately visible. Identity verification using photo identification prevents impersonation, while requiring test takers to sign attestations acknowledging rules and consequences creates both legal documentation and psychological deterrence. Clear oral instructions before the test begins should emphasise prohibited behaviours and their consequences.

Finally, multiple test versions represent an effective strategy for reducing collusion opportunities. By printing alternate forms with items or sections presented in different sequences, test providers ensure that adjacent test takers work with different materials, making copying less feasible. This approach requires careful validation to ensure versions are equivalent in difficulty, but the investment pays dividends in fraud prevention. Colour-coded test booklets or headers can help invigilators quickly verify that test takers are working on appropriate versions.

4.6.2 Digital test fraud prevention and detection

In digital testing, platform security features provide the technological backbone for fraud prevention. Lockdown browsers prevent test takers from accessing unauthorised applications, websites, or files during the examination, essentially creating a controlled digital environment. Screen recording capabilities document every action taken during the test session, while keystroke logging and answer pattern analysis can flag suspicious behaviours such as unusually rapid responses or copy-paste actions. Randomisation algorithms that present items in different orders for each test taker or draw from large item pools (see on-the-fly item selection) make it significantly more difficult for test takers to share information about specific questions.

Authentication and monitoring technologies verify identity and maintain test integrity throughout the session. Biometric verification through facial recognition, fingerprint scanning, or continuous authentication that periodically confirms the test taker's identity prevents impersonation and ensures the registered test taker remains present.

For remote testing, webcam monitoring, whether through live proctoring or AI-powered systems, can detect unauthorised persons in the testing environment, prohibited materials within camera range, or suspicious behaviours such as looking away from the screen for extended periods. These systems should be configured to balance security with privacy concerns and minimise false positives that disrupt legitimate test takers.

Environmental controls extend fraud prevention beyond the immediate digital interface. For tests administered in controlled centres, workstation positioning should prevent test takers from viewing each other's screens, with privacy screens or physical barriers providing additional protection. Network restrictions can block external communication channels while permitting necessary test platform connections. For remote testing, environmental scans before test initiation, requiring test takers to use webcams or smartphones to show their surroundings, verify that unauthorised materials or persons are not present. Clear policies should specify permissible room conditions, such as closed doors, absence of other people, and cleared desk surfaces.

Behaviour analytics, although **detection** rather than **prevention**, leverage the unique capabilities of digital systems to identify anomalous patterns. Response time analysis can flag answers provided unusually quickly, suggesting pre-knowledge of questions or external assistance, or unusually slowly in ways inconsistent with normal test-taking behaviour. Pattern recognition algorithms can detect similarities across test taker responses that exceed statistical probability, suggesting collusion or shared answer sources. Navigation patterns, such as repeatedly leaving and returning to items or accessing items in suspicious sequences, may indicate consultation of external resources during brief disconnections or periods when monitoring is circumvented.

4.6.3 Cross-modal fraud prevention and detection strategies

Clear communication serves as a foundational deterrent across both delivery modes. Test takers must receive explicit information about prohibited behaviours, security measures in place, and consequences of violations before registering and again before test administration. (See Section 4.3.1.) Visible reminders during testing, whether through invigilator presence or on-screen warnings, reinforce expectations. Transparent policies regarding how suspected fraud is investigated, what evidence is required for sanctions, and test takers' rights to respond to allegations ensure procedural fairness while demonstrating serious institutional commitment to integrity.

Staff training ensures that those responsible for test administration can effectively implement fraud prevention measures. Both invigilators and remote proctors need regular updates on emerging fraud techniques, recognition of subtle indicators, appropriate intervention protocols, and documentation requirements when incidents occur. Technical staff supporting digital testing must understand platform security features, troubleshoot without compromising test integrity, and recognise data patterns suggesting fraudulent activity. Training should emphasize that prevention serves all test takers by protecting the value and credibility of their legitimately earned results.

Post-test analysis provides essential verification that extends beyond real-time monitoring. Statistical analysis can identify unusual score distributions, unexpected performance patterns, or

correlations between test takers suggesting collusion. For paper tests, comparison of handwriting across different test sections or against registration signatures may reveal impersonation. For digital tests, review of recorded sessions can confirm or refute real-time alerts from automated systems. These analyses should follow documented protocols that distinguish between genuine fraud and legitimate variations in test taker performance, ensuring that sanctions are applied fairly and only when evidence is compelling.

4.6.4 The primacy of prevention over detection

Prevention fundamentally outweighs detection in maintaining language test integrity because it protects the validity of the testing system proactively rather than responding to damage already done.

Once fraud occurs and goes undetected during administration, contaminated results enter the test taker pool, distorting norm-referenced scoring, compromising item calibration data, and potentially enabling fraudulent test takers to gain undeserved certifications with real-world consequences.

Post-test detection, while valuable, can only retrospectively cancel scores or revoke certificates, but cannot undo the statistical contamination of psychometric analyses, recover costs associated with investigations, or fully restore stakeholder confidence once breaches become public. Moreover, effective prevention creates a culture of integrity that deters potential cheaters before they act, whereas detection relies on catching violators after the fact, often requiring resource-intensive investigations, complex evidence gathering, and difficult adjudication processes that burden both test providers and innocent test takers who may face delays or additional scrutiny.

Security resources should be allocated hierarchically: prevention measures provide the highest return on investment by stopping fraud before it occurs. Test programmes should ensure robust prevention systems (secure environments, trained proctors, identity verification) before investing heavily in detection technologies. Behaviour analytics and statistical detection methods serve as secondary safeguards, not primary security measures.

The extensive discussion of detection methods in this chapter reflects their technical complexity, not their priority.

4.6.5 The importance of deterrence strategies

Deterrence strategies prove equally important as prevention or detection mechanisms. Clear communication before and during test administration about prohibited behaviours, potential consequences, and monitoring capabilities helps establish behavioural expectations. Visible security measures, whether invigilator presence or on-screen reminders about proctoring, create psychological deterrence while demonstrating the test provider's commitment to fairness.

In-test alerts when test takers engage in unauthorised actions (e.g. changing tabs in the browser or opening other application windows) usually have a very discouraging and immediate effect.

Regular audits of both paper and digital test sessions, including random reviews of recordings and answer patterns, maintain ongoing vigilance while identifying emerging fraud techniques that require countermeasures.

4.7 Key questions

- Will delivery be on paper or digital? Or both? If digital, what types of devices will be used (computers, tablets, phones, etc.)?
- Are tests offered remotely, with proctoring?
- What procedures and criteria are in place for selecting the testing centres/ the testing platforms?
- If digital, how do test delivery settings affect item writing and test construction?

- What resources are available to administer the test (administrative staff, invigilators, rooms, audio, etc.)?
- How will test takers with special needs be accommodated?
- How should staff be trained?
- How can resources such as rooms and devices be checked before the day of the exam?
- How can test takers try out or practise a digital test?
- How will test takers be registered and their attendance at the exam be recorded? How will test takers be identified?
- How will materials be transferred to and from the venue(s)?
- How will test materials be securely stored before administration? And after administration?
- What could go wrong? Are there procedures and regulations about what to do if this happens?

4.8 Further reading

See ALTE (2001m) for a self-assessment checklist for logistics and administration.

See Dirkx et al. (2014) for the testing effect on learning.

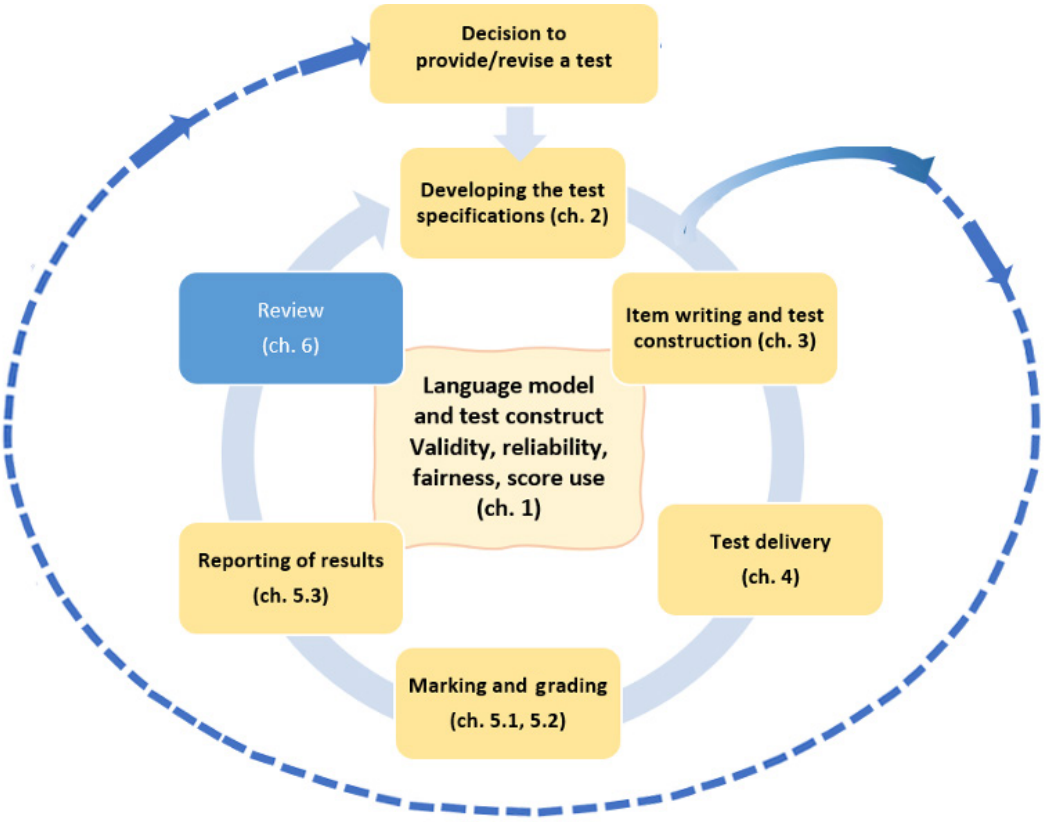
See Tsagari and Spanoudis (2013) and Taylor and Banerjee (2023) for assessing L2 students with special needs.

Next steps

Chapter 4 focused on the actual test delivery, or from the test taker's perspective, the test taking. It is primarily a matter of operational organisation and logistics, rather than content. However, the formats in which a test is taken largely determine the test taker's chances of proving themselves to the fullest. Technology is playing an increasingly important role in this. That role is primarily facilitative and enabling, but should never be an obstacle to equal opportunities for test takers.

After the test delivery, the data collected during the test must be processed; the test takers' answers must be interpreted, evaluated, scored, analysed and transformed into a concrete result, report or certificate. This will be the subject of **Chapter 5**.

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 (score) 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 takers, such as whether to be accepted as part of a job application. Figure 10 gives an overview of a common version of the process. In some cases, however, the evaluation of test taker performance (marking) 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.

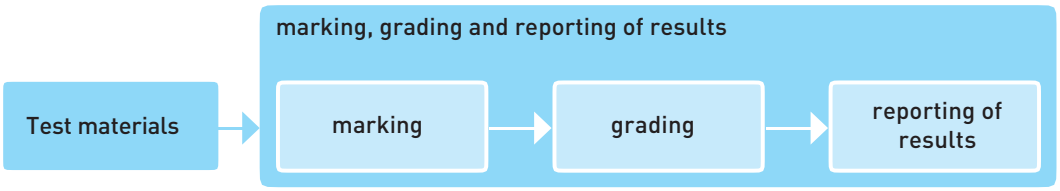


Figure 10 ***The process of marking, grading and reporting of results***

Preliminary steps

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

- Develop the approach to marking and grading (see Section 2.4.4)
- Recruit markers and raters
- Train markers and raters

As in many other chapters, technological evolution has left an undeniable imprint on the process of marking, rating, grading and reporting of results. Technology made its entry years ago as a support tool for making this process more efficient; digital item banking and digital delivery of language tests provide possibilities that are unimaginable with paper tests, such as automated marking of closed questions with immediate availability of results and reports, but also streamlining of the evaluation process of open-ended questions, where answers are immediately available digitally in a platform and thus no scan-and-distribute-to-raters procedure is needed. Technology can be used for automatic detection of key words or key phrases in open-ended answers or for flagging language errors (e.g. spelling, grammar) in text answers to assist the human evaluator; artificial intelligence (AI) may also be used to convert spoken language answers into text, search for similarities with a model answer, detect plagiarism, evaluate responses according to rating criteria, and more. These advances have been enabled in part by the greater ease with which corpora of test takers' written and spoken production can be built up, structured and analysed using AI tools.

Even though the methods of marking, rating, and grading have expanded, the basic principles of ensuring that the marks are appropriate and reliable are the same regardless of the mechanism: test providers are responsible for specifying mark schemes, rating scales, training (of humans and/or machines), quality assurance, and accurate reporting (see Sections 1.5.2 and Appendix III). Human judgement remains the basis of ratings, whether mediated by corpora and algorithms, or honed by careful training.

This chapter focuses on those principles.

5.1 Marking

While the expression *marking* covers all activities by which marks are assigned to test responses, a distinction needs to be made between clerical marking, marking of short answers, and rating of written and oral productions. All of these can be performed by humans or by automated systems.

5.1.1 Clerical marking

Clerical marking involves matching the test taker's responses to the answer key, with no variation possible or permitted. It applies to types of items like binary or multiple choice, in accordance with answer keys which need to be accurate and clear.

5.1.1.1 When done by humans

The markers need to be attentive to detail, prepared to engage in repetitive activity and careful when calculating the points. When spotting wrong answers recurring in multiple papers, they should flag the key so that it can be checked for accuracy, since this might point to an error in the answer key. If an error is spotted, the key needs to be corrected by the exam developer, the corresponding test takers must be credited with the points, and all or part of the exams need to be checked in relation to that particular item. Markers can consult with each other and with the exam developer to make sure that all results are accurate.

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.

Training would involve familiarisation with the procedures to follow and with the mark scheme on a regular basis.

Permanent monitoring of the quality of work needs to be ensured. Markers' work should be recorded and analysed regularly, and action can be taken following the results (e.g. closer monitoring if needed).

5.1.1.2 When done by automated systems

Human expertise needs to be in control and periodic checks on the quality of marking should be carried out. During post-test analysis (see Appendix III), items that are much more difficult than expected or that discriminate poorly should be checked to ensure that the key was properly identified. Ideally, this will take place before scores are reported. Typically, if pretesting is used, such errors can be caught before the items in question form part of the official score, and at the point where the item scores must be reported, errors have already been dealt with.

5.1.2 Marking of short answers

For tasks with short answers, test takers typically write one or a few words. For such items it may be difficult to ensure that the key is exhaustive. The marking process is thus more complex than clerical marking, as some judgement is required.

5.1.2.1 When done by humans

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. If the range is determined relatively objectively (perhaps one mark may be awarded for selecting a suitable verb and a further mark for using the correct form), a list of possible responses and points may be sufficient; however, if the degree of success involves words that are more or less appropriate for the context, some level of relevant expertise is needed for markers to distinguish between a fully correct, a partially correct, and an incorrect response.

It is helpful if the markers can identify and report alternative responses from test takers. The markers need to verify if these answers might be valid for the item and collect them. The answers which prove to be right need to be included in the answer key and test takers must be credited for them. All or part of the exam might need to be checked in relation to the items for which alternative responses were accepted.

Training might involve familiarisation of the markers with the procedures and training with alternative responses given by test takers.

For monitoring the quality of work, an evaluation system based on a number of parameters, such as accuracy, reliability and speed, is useful. As there is a subjective element involved, double marking will increase the reliability of the results. Action can be taken based on results (e.g. retraining or revision of the mark scheme).

5.1.2.2 When done by automated systems

Procedure without AI

Short answers which are marked as incorrect by the system need to be checked by human markers in case they might be acceptable. All answers which might be considered acceptable need to be indicated as such to the automated system and if needed exams will be reassessed.

Procedure with AI

Where AI, a Large Language Model (LLM) or a deep learning procedure is used, the language tester will need to explain how the reliability and accuracy of the model have been implemented and the extent to which it is a robust procedure. The procedure will be accompanied by an evaluation of the level of certainty of the assessment produced by the model and a proposed procedure for validating that assessment and, where needed, explaining outcomes (e.g. criteria, linguistic range, diagnostic function). The procedure for validating the machine ratings should be commensurate with the use that will be made of the results. In all cases, the language tester developing the test is expected to provide evidence and guarantees of correct operation and consequently of reliable marking.

5.1.2.3 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 to employ more markers than estimated, to ensure that any difficulties can be coped with.

If an automated marking system is used, necessary time for marking should not be an issue. However, the quality of the marking process needs to be constantly monitored. As with clerical marking, errors may be caught by post-exam analysis; in order for the errors to be corrected in a timely way, at least a provisional post-exam analysis needs to be carried out as soon as the marking process is complete, but before the scores are reported.

Marking via an online platform needs to run through secure systems for exam confidentiality and protection of personal data.

5.2 Rating

Rating refers to the assessment of oral and written production. We will use *rating* and *raters* here to refer to marking where it is necessary to exercise trained judgement to a much greater degree than in clerical marking. In this case, 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, both 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 and the marking of short answers is also true of rating: the process should be managed to use resources effectively, and checks and monitoring should help ensure accuracy. The reliability of ratings should also be monitored (see Appendix III). However, the rating of oral and written production is more complex since it is not a matter of all or nothing, that is, the production of a test taker is not either correct or incorrect and it may display different levels of achievement. The oral and written performance of test takers is made up of different dimensions within which there are also various levels of proficiency. In the assessment of oral and written performance three aspects dovetail, namely the performance of test takers, the reliability and accuracy of the work done by raters, and the reliability and appropriateness of the tool that the latter uses to analyse the former. Out of these three elements, the only one that we cannot control is the performance of test takers. The other two (the reliability and accuracy of the work done by raters and the reliability and appropriateness of rating scales) must be calibrated as carefully as possible. For a discussion of the development of rating scales, see Section 2.4.4. The reliability of raters' work is discussed below.

5.2.1 The rating process

For rating to work well, raters – and/or automated systems – must have a shared understanding of the standard. The basis of this shared understanding are shared examples of performance.

For small-scale exams, a group of raters may arrive at a shared understanding through free discussion. This may mean that raters are in good agreement, 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 documentation and a standard training process to communicate the standard to newcomers, as well as methods to ensure that all raters are within a reasonable degree of agreement. The degree of agreement for such large-scale exams can be evaluated by a number of statistical properties, for example inter-rater reliability coefficients, intra-class correlation coefficients, and Many-facet Rasch Measurement (MFRM). See Appendix III for details on these procedures.

5.2.2 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 (*Manual for Relating Examinations to the CEFR* and *Aligning Language Education with the CEFR*). It may also be necessary to 'train raters out of' using a rating scale which they are already familiar with, if that scale is not compatible with the standard being implemented. 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 and the marking scale
- 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

If automated rating is used, typically the algorithm will require numerous human-rated samples at each level, and the test developers will evaluate the accuracy, reliability, robustness and explainability of the algorithm. The way this is done will depend on the nature of the automated rating system and the use of the scores or levels.

If the rating scales have been designed as task-specific (see Section 2.4.4), rater training (for humans or machines) will have to be undertaken for each new prompt or task. If the rating scales are generic, training should be timed so that there is not a large gap between training and live rating. If tests are administered only once or twice a year, training should be repeated before each administration. If tests are administered frequently, human raters should be retrained periodically to ensure that the rating standard is being maintained.

5.2.3 Monitoring and quality control

Ideally the training phase ends with all raters working sufficiently accurately and consistently to make further feedback or correction unnecessary. This allows the rating phase to continue as smoothly as possible.

However, some monitoring is necessary to identify problems early enough to remedy them.

Whatever way the rating is done, there must be a way to ensure that the rating meets the required level of quality. Monitoring (See Chapter 6) will help account for the validity and the reliability of the scores and grades to be delivered.

Regarding validity, it is important to ensure that, when making their judgement, the raters or the algorithm take into consideration the different aspects of the production that were embedded in the construct during the design of the construct, the rating guidelines or the rating scale (e.g. rating criteria). They should base their judgement only on relevant features, not introducing aspects external to the construct. They should also have an adequate representation of what is expected from a test taker at each level targeted by the test.

A way of observing the validity of human judgements is to run qualitative studies (e.g. using verbal protocols as a diagnosis tool; see Chapter 3, especially Section 3.3.3.3), but it cannot be done on a large scale. Analysis of live test data and ad hoc quantitative studies will usually provide more possibilities to monitor each rater or the rating algorithm. Such a profiling will enable the provision of individual feedback or training to improve the quality of the rating or, for instance, to identify the need for revision in the rating scale.

Here are some kinds of problematic 'rater effects' that can be identified (these effects also apply to rating algorithms):

- **Central tendency**. The rater's range of marks is too narrow and therefore does not distinguish high and low performance clearly enough
- **Severity or lenience**. The rater's marks are systematically too low or too high
- **Halo effect**. A rater who has to assign marks on several criteria forms an impression of the test taker when awarding the first mark, and applies it to all the other marks, irrespective of the actual level of performance

Such effects, which are a threat to validity, also have a negative effect on the agreement between raters, i.e. on inter-rater reliability. When such an effect is systematic for a given rater, it can easily be identified, and action can be taken to reduce its impact on scoring or to change the rater's (or algorithm's) behaviour.

The corrections which are possible correlate with the kinds of problems which are identified. Regarding severity, it appears that many raters may have an inbuilt level of severity and attempts to standardise this out of existence can backfire, by reducing the markers' confidence and making their judgement more erratic. Thus, it may be preferable, depending on the testing context, 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 III). The use of too narrow a mark range can only be partially corrected statistically, and there is no way to compensate for the halo effect, so both problems need to be identified and remedied, mainly by retraining the raters.

Another problem is that human raters may be inconsistent or even erratic in their rating, which cannot be predicted and will reflect badly on their intra-rater reliability and, as a consequence, on inter-rater reliability. Inconsistent rating is impossible to correct statistically and may jeopardise statistical adjustments made to deal with rater effects. Continuous training may help to reduce the lack of consistency in rater judgements.

Some system of monitoring and adjusting is thus desirable. It may be implemented more easily for writing, where a script may easily be (re)assessed by several raters and also by an algorithm. Such monitoring for oral assessment requires recording the test taker's performance. If such a recording is not possible, 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 III).

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 written responses in a marker's allocation and observe how closely their marks agree. However, for this to work reliably these responses should be indistinguishable from others, so photocopying, for example, is not possible. Practically this method is feasible only for written production elicited in a digital test, or where paper scripts are scanned or (for speaking) recordings are uploaded into an online marking system.

Another way of reducing rating error, and of comparing raters with each other (which helps to identify and correct some rater effects) is to use double rating or *partial multiple rating*, where a proportion of scripts or recordings are rated by more than one rater, or parallel rating by an algorithm. Depending on the statistical approach used, some method is needed to combine the information and arrive at a final mark for the test taker (for example averaging the marks, the experienced rater overriding the less experienced rater, providing a third marking if the two marks are not within tolerance, etc.). Multiple rating should be used for statistical analysis and for quality control in live sessions through flagging of discrepant results. Figure 11 gives a workflow example for rating an oral exam using a digital platform, showing several rating 'rounds' with their respective characteristics: live rating; followed by a post-test rating round by an external rater; followed by an automatic decision round where the system calculates the average of the previous rounds and automatically flags scores that are more than 20% apart. In this example, productions of test takers are flagged when there is a certain level of disagreement between the live rating and a second rating (be it by a human or an algorithm) based on the recordings. Such flagged productions must be checked before delivering a final score to the test takers.

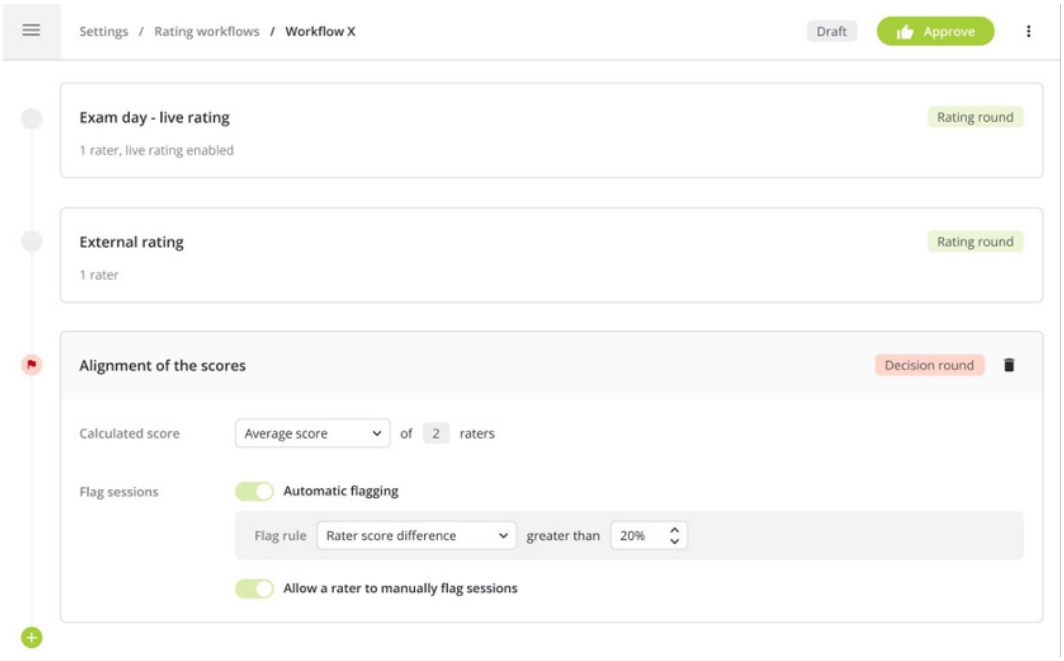


Figure 11 ***Example of a technology-supported rating workflow for an oral exam***

5.3 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 levels (CEFR or other proficiency scales). See Section 2.4.4 for design considerations in this area.

Taking as an example the CEFR, 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 defined as 'passed' and the others as 'failed'. Degrees of passing or failing may also be indicated. If several CEFR levels are considered, the pass or fail marks have to be differentiated accordingly.

Identifying the score which corresponds to achieving a certain level is called *standard setting*. It inevitably involves subjective judgement which should be based on evidence as far as possible.

Different approaches to standard setting apply to tasks testing language production and to items and tasks testing comprehension, which are often objectively marked. Standards for tests of production can be set in a more intuitive way, as a written or oral performance represents real-life competence more directly than responses to tasks targeting comprehension.

Standards for tests of comprehension are harder to set, because we must 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 components that are scored separately, a standard must be set for each of them separately, leaving the problem of how to summarise the results (see Section 5.4 for

more on this). The reader is referred to the *Manual for Relating Examinations to the CEFR* and *Aligning Language Education with the CEFR*, which contain extensive chapters on standard setting.

Strictly speaking, standard *setting* is something that should happen once, when the exam is first developed. Over time, however, the process should be considered 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 for Relating Examinations to the CEFR* (North & Jones, 2009).

5.4 Reporting of results

The reporting of results should not be seen as merely communicating a test taker's result or score. Instead, results reporting should reflect the findings about the skills or competencies tested and the test objectives (see Section 2.6.2.). Test results must be understandable and easy to interpret by all stakeholders. It should be clear for test takers, teachers, employers, authorities, etc. what level of proficiency is certified, how this decision was reached, what sub-scores determined the final result, etc.

The raw score resulting from adding up all the marks from a given component is usually not as helpful as a scaled score (scaling of raw scores enables them to be interpreted more consistently across test versions) or level designation. Beyond the mere score and result itself, all stakeholders must understand the value of the results: what they can be used for and where and how the authenticity can be checked. The value of the results is clearly also important from a regulatory perspective. Reports need to be formulated in such a way that they are readily understood by the test takers and those stakeholders (score users) who ask for certificates to be handed in as a basis for decisions, for employment, university entrance, migration and other purposes.

The use of an established framework such as the CEFR will certainly contribute to such interpretability. The CEFR Companion volume stresses the importance of reporting results as a profile of proficiency with respect to language activities relevant to the score use situation, where possible, as opposed to a single level result across all activities and modes of communication. It is important to acknowledge, however, that some stakeholders require a single result aggregated across all test components. Where a single result is required, a method must be chosen for aggregating the marks for each test component. The testing organisation must decide how each component should be weighted – all of them equally, or some more than others – and should take care to document the rationale for this, taking into account (beyond external requirements) the way in which the result reflects the dimensions of the construct. This may require some adjustment of the calculation of raw or scaled scores on each component. (See Section 2.4.4.)

The use of technology can greatly contribute to the visualisation of results. For example, it can allow clicking through an overview report with visual elements supporting better and faster result interpretation (Figure 12). It can also show increasingly detailed underlying scores and results. In Figure 13, each row represents the scores per item for one test taker. The report allows seeing correct responses (in green), incorrect responses (in red), unanswered items (in grey), partially correct responses (in orange), and neutralised responses (removed from scoring after item analysis).

Finally, technology is also an important tool to make each results report unique, identifiable and traceable. This gives all stakeholders (test takers, issuing authorities and score users) the guarantee that the results are valid, fraud-free and consistent. High-stakes language test reports or certificates should have a unique and verifiable identifier, allowing for public verification of authenticity.

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

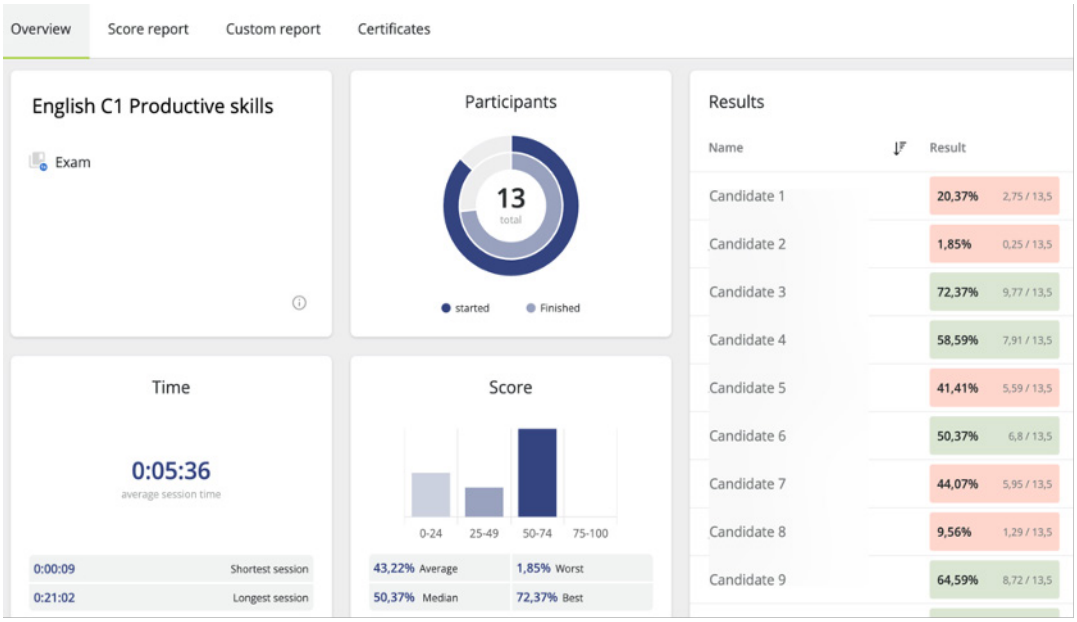


Figure 12 ***Overview test report***

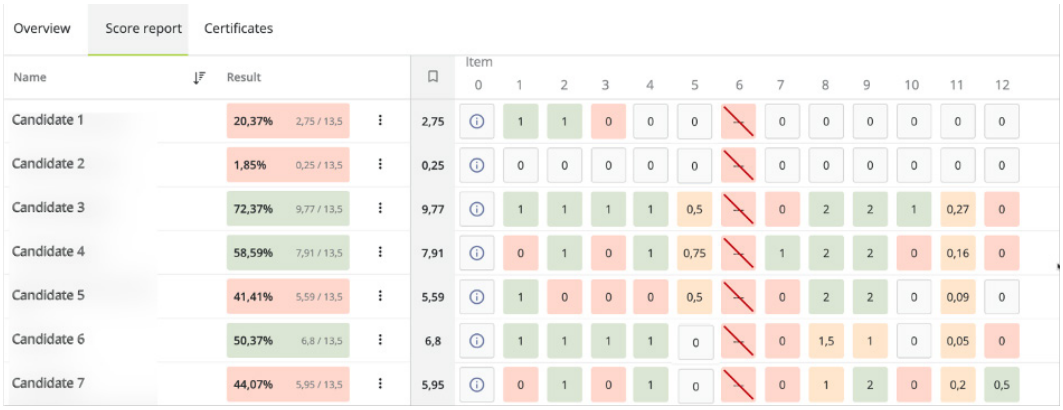


Figure 13 ***Detailed item score report***

5.5 Key questions

- How much clerical marking will your test require and how often?
  - Will there be short-answer items, and how will they be marked?
  - What procedure for modifying an answer key during the marking process is in place?
- 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 enough?
- What is the best way to grade test takers in your context?
- Who will you report results to and how will you do it?

5.6 Further reading

See ALTE (2001n) for a self-assessment checklist for marking, grading and reporting results.

See *Manual for Relating Examinations to the CEFR* and *Aligning Language Education with the CEFR* for procedures of standardisation and standard setting. Kaftandjieva (2004), North and Jones (2009), and Figueras and Noijons (2009) also provide information on standard setting.

Weigle (2002) presents scoring procedures for assessing writing.

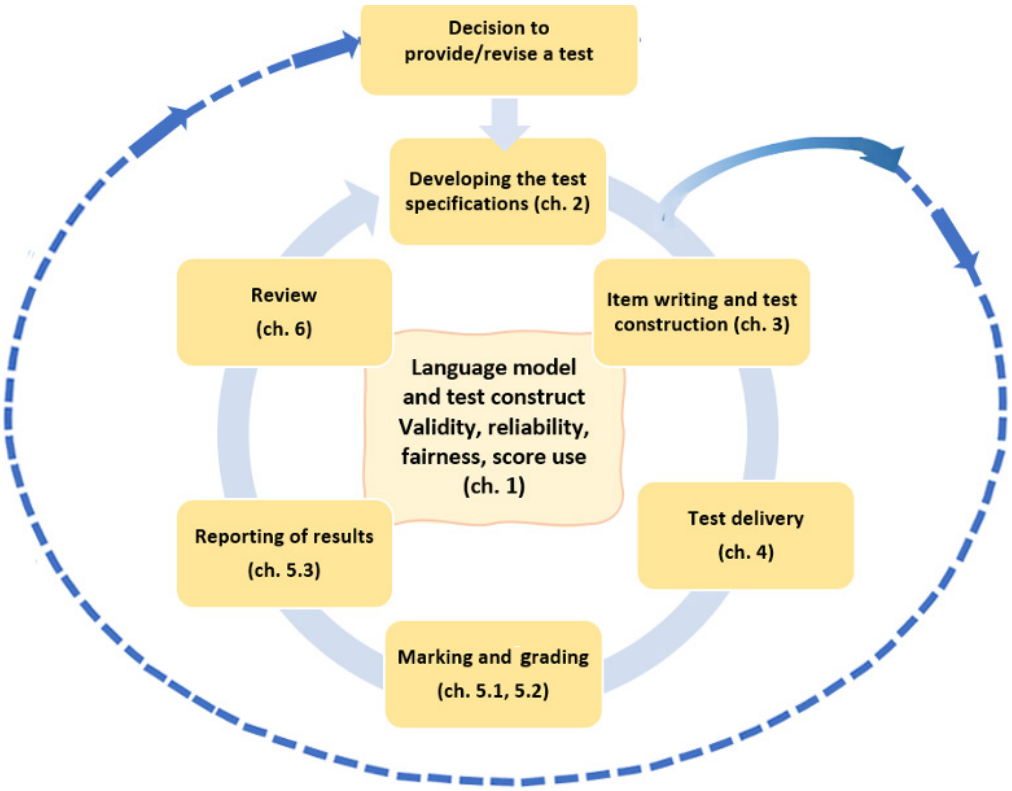
Shermis and Wilson (2024) provide information about automated essay evaluation.

Next steps

Chapter 5 addressed the work required to translate the information from test takers' individual responses into information that stakeholders can use to make decisions, from the process of giving marks or ratings to responses, through the process of aligning those marks or ratings to the CEFR or another grading standard (standard setting), to considerations of reporting the results to stakeholders.

Chapter 6 will explore review, monitoring, and revision as essential components of a validity argument. It will consider information available from all stages of the Testing Cycle and will provide guidance for using that information to ensure that the test is functioning as intended.

6 Monitoring, review and revision



It is important to conduct regular checks of the work done to develop and administer the test. Is it of an acceptable quality, 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 in coming sessions.

Test review is a kind of project which looks at many aspects of the test. It also goes back to the decision to provide a test and 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.

6.1 Routine monitoring

Monitoring is part of the routine operation of producing and using the test.

Monitoring refers to observing and checking the processes of test development and administration on an ongoing basis. It ensures that the test functions as intended and remains responsive to stakeholder needs. (See Section 1.6.) Monitoring can take many forms:

- Observing and analysing test taker responses and results
- Gathering and interpreting feedback from test takers, raters, or administrators
- Conducting statistical analyses of item and test performance
- Reviewing marking and grading processes
- Assessing the impact of the test on test takers, teachers, and institutions

The aim of monitoring is to verify that important aspects of a test are acceptable during or soon after its use. Because monitoring happens in real time or shortly after administration, it allows test providers to identify and address issues quickly, often improving the experience for current or future test takers. These can refer to preparation or content of test materials, test delivery conditions, meeting specific statistical targets, etc.

Evidence gathered for monitoring will be used to ensure that everything concerned with the current test version is on track: that materials are properly constructed, that they are delivered on time, that 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 Chapter 1 and Appendix I), and should also be reviewed with this in mind.

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

- Using expert judgement and trialling or pretesting to ensure items are well written (see Section 3.3.3)
- Using test taker responses to decide if items work appropriately (see Section 3.3.3, and also Appendix III on detecting bias using Differential Item Functioning)
- Using feedback forms to see if administration went well
- Collecting and analysing data about human and automated marking, and rating performance (see Chapter 5 and Appendix III)

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

Review is broader in scope. It involves a systematic evaluation of many aspects of the test and may revisit fundamental questions of test purpose, construct, and use. (See Section 1.6.3.) Review cycles can occur periodically and be scheduled checks (per administration, quarterly, annually, every five years, etc.) with clear owners and deadlines, drawing on data and experience gathered over time. A review cycle may be carried out 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.

When carrying out a periodic review, test providers should look beyond routine monitoring data and bring together evidence from several sources. Useful elements include:

- **Stakeholder feedback**. Collate and interpret input from test takers, institutions, and appeals to identify recurring themes or concerns. Stakeholder engagement ensures that the perspectives of test takers, teachers, institutions, and other users of the test are systematically collected and considered. This can be done through surveys, focus groups, structured feedback forms, or consultation meetings. Quantitative studies can be carried out to evaluate how well the test predicts future success or correlates with other measures. Effective engagement helps identify issues early, improves transparency, and builds trust in the assessment system. Appeals procedures provide test takers and institutions with a formal channel to contest results or administrative decisions. Clear, accessible policies are essential to maintain fairness, protect test takers' rights, and demonstrate institutional accountability. All these procedures should be reviewed from time to time to ensure their continued suitability and effectiveness.
- **Performance indicators**. Compare outcomes with expected benchmarks, for example reliability targets, or with impact metrics such as test taker outcomes, paying attention to unusual results

and patterns that may indicate systemic issues (see Chapter 5 on statistical monitoring and Appendix III for types of statistical indices).

- **External review and accountability.** Independent evaluation of a testing programme by external experts, partner institutions, or accrediting bodies. It provides a broader perspective, validates internal findings, and helps align the programme with sector-wide standards.
- Accountability means that test providers can demonstrate to stakeholders that their assessments are fair, valid, and reliable. This involves publishing policies, reporting results transparently, and documenting the rationale for decisions. Together, external review and accountability reinforce the credibility of the test. (See Sections 1.4, 1.6.1, and Appendix I)
- **Governance and quality assurance frameworks**. Test providers should have in place structures and processes that define roles, responsibilities, and decision-making in test development and delivery. Clear governance ensures that responsibilities for monitoring, review, and improvement are assigned and carried out effectively.
- Quality assurance frameworks provide the systematic procedures and tools used to plan, check, and improve the quality of assessments. This may include checklists, validation studies, rater training, and scheduled reviews. A well-defined framework links evidence from monitoring and review to decision-making, ensuring continuous improvement. (See Appendix I for the ALTE Minimum Standards framework.)
- **Institutional planning**. Resourcing decisions should be linked to broader institutional planning. This means aligning budgets, staffing, and professional development with the organisation's long-term objectives for test quality and sustainability. Institutional planning ensures that resources for test development, monitoring, and review are not only available in the short term but are also embedded in annual cycles and strategic plans, supporting continuous improvement and stability over time.

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 revisions or no changes.

The diagram of the Testing Cycle highlights the periodic review described in Section 6.2, plus permanent monitoring (see Section 6.1). It shows that the findings of the periodic review feed into the very first stage in the diagram: the decision to provide a test. The test development process may be completed again as part of the review. The permanent monitoring activities, however, are conducted at any point in the cycle and may prompt changes at any point in the cycle. Care should be taken to consider the impact of any revisions on the validity argument (see Section 1.4), which may need to be changed as a result, and 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. Reviews 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. See Chapter 1 and Appendix I for more information on validity arguments, and Appendix III for information on statistical analyses used for monitoring.

Other tools that may be helpful in monitoring and review include the ALTE Content Analysis Checklists (ALTE, 2001a-o). Jones et al. (2006, pp.490–492) provide a list of 31 key points for small-scale achievement testing. Another useful reference is the *Standards for Educational and Psychological Testing* (AERA et al., 2014).

6.4 Revision

Revision refers to the process of implementing changes in relation to different elements of the test, such as items, specifications, rubrics, or administrative procedures. Revision may follow from monitoring or review; once evidence has been collected and issues identified, concrete modifications are made to improve the test.

Revisions can vary in scope:

- **Minor revisions.** Small adjustments such as clarifying rubric wording, replacing an item, or correcting a technical issue
- **Moderate revisions**. Updating parts of a test specification, revising rating scales, or adjusting administration procedures
- **Major revisions**. Substantial changes such as rebalancing constructs, introducing new task types, or overhauling marking procedures

Minor revisions typically require only internal documentation and affect only one or two stages of the Testing Cycle. (All changes should be documented: see Section 1.6.2.) Moderate revisions may require communication with stakeholders to inform them of the changes, and may in some cases require a new equating or scaling; however, they typically do not substantially change the interpretation of the score. Major changes generally require changes to most or all stages of the Testing Cycle, a new standard setting, and changes in the interpretation of the score that require extensive stakeholder communication.

It is not always obvious whether a given revision is minor, moderate, or major. For example, if a change is made to the scales raters use to evaluate test taker performance for a given test component (e.g. spoken production), will that change the range of potential scores, or have a major effect on the scores that would have been received by test takers? Will a new standard setting need to be conducted? Do communications need to be prepared for test takers and score users? Will the interpretation of scores be the same as they were previously? (See Sections 2.4.4 and 5.2 on rating scales.) If the change is to clarify the interpretation of a term, and the change does not result in changed scores, the change may be limited to just the rater training documents. However, if the change is to add or remove a category of rating (for example, to add fluency as a rating factor), this will likely result in major changes in scoring, a new standard setting, and a 'breaking of the scale,' meaning that interpretations of scores before the change will not be the same as interpretations of the scores after the change.

Therefore, every revision must be carefully examined to determine the consequences of making the change versus not making the change, as any change has the potential to affect the validity argument. (See Section 1.4.) This is not to say that no revisions can be made without a major committee meeting; certain types of changes may be common enough that standard procedures can be developed, so they can be implemented smoothly and without too much disruption, whereas other types would be identified as requiring approval by different levels of the organisation or stakeholders.

Effective revision is guided by three principles:

1. **Evidence-based**. Changes should be justified by data (statistical analysis, stakeholder feedback, monitoring findings, examination of all consequences)
2. **Documented.** All revisions should be logged, with reasons, actions taken, and responsible persons recorded for accountability
3. **Cyclical.** Revision is not a one-off exercise but part of an ongoing cycle where monitoring and review generate evidence, and revision ensures continuous improvement

Revision therefore serves as the practical link between *review* (evaluation and decision-making) and *implementation* (ensuring that the test remains valid, reliable, and fair in practice).

6.5 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?
- What specific process will be used to document revisions?
- Who needs to approve different types of revisions?
- How will the stakeholders be involved in the review process?
- How will the stakeholders be informed about the changes to the test?
- How will the impact of the implemented changes be monitored?

6.6 Further reading

The ALTE Minimum Standards (2024) provides a number of categories through which to assess a test. ALTE (2001o) 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).

Conclusion

Chapter 6 concludes the Manual by showing that test quality depends on continuous monitoring, periodic review, and evidence-based revision. Routine monitoring checks that each test administration functions as intended, while broader reviews revisit the test's purpose, construct, and use in light of accumulated evidence. Revision then translates findings into practical improvements, ensuring that the test remains valid, reliable, fair, and responsive to stakeholder needs. Together, these processes close the Testing Cycle and support the long-term sustainability of a strong validity argument.

Appendix I – Building a standards-based validity argument

Standards for test quality are helpful ways to break down a validity argument (see Section 1.4) into actionable criteria to meet. The ALTE Minimum Standards (MS) are an example; they are used within ALTE's quality assurance system to define those qualities in a test that are necessary to obtain the ALTE Q-mark quality designation, but they are helpful to assess test quality even without engaging in a formal audit. Outside ALTE, other quality frameworks may be used, but a discussion of these is beyond the scope of this Manual. This Appendix presents the ALTE MS with practical applications as a framework for identifying activities to support the quality of language tests in any context.

The five aspects of test development with ALTE Minimum Standards

Aspect 1: Test Construction (MS 1–5)

This aspect focuses on how a test is designed and developed (see Chapters 1–3 and 5 in this Manual). It covers the foundations of test purpose, theoretical grounding, team expertise, test version equivalence, and alignment to external frameworks such as the CEFR. Decisions at this stage determine what the test claims to measure, how defensible those claims will be, and how much confidence stakeholders can place in the outcomes. Strong test construction practices ensure that the test is both meaningful in its intended context and sustainable across administrations.

1.1 MS 1: Test purpose and population
Standard: You can describe the purpose and context of use of the examination, and the population for which the examination is appropriate.

What this means in practice: You must provide a comprehensive description that clearly defines why your test exists, who should take it, and how the results will be used. This involves specifying the target population's demographic characteristics, linguistic backgrounds, educational levels, and motivational factors. You need to articulate the specific contexts in which the test will be administered, the frequency of administration, and the validity period of scores. Additionally, you must explain how test results will be interpreted and applied in decision-making processes, whether for placement, certification, admission, or other purposes. This description forms the foundation of your validity argument and helps stakeholders understand the appropriate and inappropriate uses of your test. Documentation should include evidence of needs analysis, stakeholder consultation, and demographic analysis of your intended test-taking population.

1.2 MS 2: Theoretical construct
Standard: The examination is based on a theoretical construct, e.g. on a model of communicative competence.

What this means in practice: You must ground your test in established language learning and assessment theory, demonstrating how your theoretical framework translates into practical test tasks and scoring procedures. This requires selecting and clearly articulating a specific model of language competence (such as Bachman and Palmer's model (1996), the CEFR action-oriented approach, or other recognised frameworks) and showing how each component of your test reflects different aspects of this model. You need to explain how your theoretical construct addresses the specific language activities and competences required in your target use domain. The alignment between theory and practice should be evident in your test specifications, task design, scoring criteria, and interpretation procedures, including standard setting. You must also demonstrate that your construct definition is current, research-based, and appropriate for your target population and intended uses.

1.3 MS 3: Test development team
Standard: You provide criteria for selection and training of test constructors, expert judges and test writers involved in test construction and development.

What this means in practice: You must establish transparent, merit-based procedures for recruiting, training, and maintaining a qualified team of test development professionals. This involves defining specific qualifications, experience requirements, and competencies needed for different roles (e.g. item writers, reviewers, technical experts). You need to implement systematic training programmes that ensure all team members understand your test construct, target population, quality standards, and development procedures. Training should address technical skills (item writing, statistical analysis), content expertise (language assessment, specific domains), and ethical considerations (fairness, bias awareness, data security). You must also establish ongoing professional development requirements, performance monitoring systems, and quality assurance measures to maintain consistent standards across all team members. Documentation should include selection criteria, training curricula, competency assessments, and performance evaluation procedures.

1.4 MS 4: Parallel versions
Standard: Parallel examinations are comparable across different administrations in terms of content, stability, consistency, and grade boundaries.

What this means in practice: You must ensure that different versions of your test are statistically and practically equivalent, so that test takers receive fair treatment regardless of which form they encounter. This requires developing detailed test specifications that define content coverage, item types, difficulty distributions, and other parameters that must be consistent across forms. You need to implement systematic procedures for test construction, including item banking (if applicable), automated test assembly (if used), expert review, and piloting, pretesting, and/or trialling. Statistical equating procedures must be employed to place all test versions on a common scale, using approaches such as common-item designs or other linking methodologies. You must monitor the equivalence of the test versions through both classical and modern test theory analyses, establishing acceptable limits for differences in difficulty, reliability, and other key statistics. Quality control procedures should include expert review for content balance, piloting, pretesting and/or trialling, and ongoing monitoring of form performance to ensure maintained equivalence over time.

1.5 MS 5: Alignment to external frameworks
Standard: If you make a claim that the examination is linked to an external reference system (e.g. the Common European Framework of Reference for Languages, CEFR), then you can provide evidence of alignment to this system.

What this means in practice: When claiming alignment to frameworks like the CEFR, you must provide empirical evidence demonstrating that your test levels correspond meaningfully to the claimed framework levels. This involves conducting rigorous standard setting studies using recognised methodologies (such as Angoff, Bookmark, or contrasting groups methods) with panels of qualified experts trained in both your test and the target framework. You need to collect multiple types of validation evidence, for example concurrent validity studies with other aligned measures, expert judgement studies of test taker performances, and analysis of test taker self-assessments. The alignment process should be systematic and well-documented, involving initial specification development based on framework descriptors, comprehensive expert training, multiple rounds of standard setting with empirical feedback, and ongoing validation studies. You must also establish procedures for maintaining alignment over time, including regular review of cut-off scores, collection of ongoing validity evidence, and adjustment procedures when necessary. If you want to align an existing test to the CEFR, we recommend consulting the handbook *Aligning language education to the CEFR* (2022) and the *Manual for Relating Language Examinations to the CEFR* (2009).

Aspect 2: Administration & Logistics (MS 6–10)

This aspect addresses the systems, processes, and resources required to deliver a test fairly, securely, and consistently (see Chapter 4 in this Manual). It includes test centre management, secure handling of materials, administration support, data protection, and accommodations for test takers with special needs. Without rigorous logistics and administrative controls, even a

well-constructed test risks producing results that are inconsistent, unfair, or insecure. This aspect therefore translates the design of the test into a reliable operational reality for all test takers.

2.1 MS 6: Test centre management
Standard: All centres are selected to administer your examination according to clear, transparent, established procedures, and have access to regulations about how to do so.

What this means in practice: You must develop and implement comprehensive systems for selecting, training, monitoring, and supporting test centres and other test delivery partners to ensure standardised delivery across all locations, including online administration. This involves establishing clear criteria for approval, including facilities (space, equipment, security), staffing qualifications, and institutional capacity. Detailed application and approval processes should include site inspections, reference checks, and trial administrations where needed. You must provide training for administrators covering delivery, security, test taker management, and emergency procedures. Supporting materials should include administration manuals and ongoing help through central or regional teams. Monitoring should include unannounced visits, performance reviews, and corrective action procedures.

2.2 MS 7: Secure test delivery
Standard: Examination papers are delivered in excellent condition and by secure means; your examination administration system provides for secure and traceable handling of all examination documents, and confidentiality of all system procedures can be guaranteed.

What this means in practice: You must implement robust security measures covering the full chain of custody from creation to disposal. This includes secure facilities for production, storage, and distribution; strict access controls; and complete audit trails. For digital delivery, cybersecurity measures must cover encryption, authentication, system monitoring, and intrusion detection. For physical materials, secure packaging, tracked distribution, chain of custody documentation, and secure destruction procedures are required. All personnel with access to materials must receive security training, and independent security audits should be conducted regularly.

2.3 MS 8: Administration support systems
Standard: The examination administration system has appropriate support systems (e.g. phone hotline, web services).

What this means in practice: You must provide reliable and accessible support services for all stakeholders throughout the testing process. This involves multiple contact channels (phone, email, web, chat) with clear service standards. Knowledge bases and FAQs should be available for common issues. Staff must be trained to provide consistent, accurate responses, and quality must be monitored via call reviews, surveys, and feedback loops. Escalation processes must ensure critical incidents are resolved quickly, especially during test administrations.

2.4 MS 9: Data protection and results security
Standard: You adequately protect the security and confidentiality of results and certificates, and data relating to them, in line with current data protection legislation, and test takers are informed of their rights to access this data.

What this means in practice: You must have a data protection framework compliant with relevant legislation (e.g. GDPR). This includes privacy-by-design, impact assessments, and appointment of data protection officers where required. You must define clearly what data is collected, how it is used, and retention periods, and communicate this transparently to test takers. Security measures should include encryption, access controls, and secure backups. Systems must support test taker rights (access, rectification, erasure, portability). You must also have breach detection and incident response procedures, including required notifications.

2.5 MS 10: Special needs accommodation
Standard: The examination system provides support for test takers with special needs.

What this means in practice: You must provide equitable access to assessment for test takers with disabilities or temporary conditions. Clear policies and procedures for requesting, reviewing, and implementing accommodations must be in place. Accommodations should be based on evidence, remove barriers without compromising construct validity, and be applied consistently. Staff must be

trained in sensitive handling of requests. Accommodations should be monitored for effectiveness, including through score comparability research, test-taker feedback, and fairness reviews.

Aspect 3: Marking & Grading (MS 11–12)

This aspect ensures that test-taker performances are evaluated accurately, consistently, and fairly (see Chapter 5 and Appendix III in this Manual). It encompasses the reliability of marking procedures, the monitoring of raters, and the systems that maintain quality over time. Marking and grading are central to the credibility of test results: if scoring is not dependable, no amount of strong design or secure administration can guarantee meaningful outcomes. Establishing robust marking and monitoring processes therefore provides one of the most direct safeguards of test validity.

3.1 MS 11: Marking reliability
Standard: Marking is sufficiently accurate and reliable for purpose and type of examination.

What this means in practice: You must design marking processes that ensure consistent results. For objective items, automated scoring should be monitored with quality checks. For constructed responses, you must provide detailed mark schemes, extensive training, and calibration exercises. Ongoing monitoring (double marking/double rating, statistical analysis, re-standardisation) is required. Reliability must be measured and reported using appropriate metrics (inter-rater, intra-rater, internal consistency), with established thresholds for acceptable performance. If automated or machine-assisted scoring is used, you must provide evidence that its reliability, validity, and fairness are equivalent to human marking. This includes benchmarking against human ratings, monitoring model stability over time, and documenting procedures for investigation and override where discrepancies occur.

3.2 MS 12: Rater monitoring systems
Standard: You can document and explain how reliability is estimated and how data regarding raters of writing and speaking performances is collected and analysed.

What this means in practice: You must implement systems to monitor rater performance continuously, including the performance of automated rating systems. Data should include agreement with benchmarks, severity/leniency patterns, and consistency over time. Analytical methods may include Many-facet Rasch Measurement (MFRM) or other suitable approaches. Raters must receive regular feedback on their performance, and intervention procedures (extra training, reduced load, suspension) must be defined for underperforming raters. Documentation must show the standards applied and the consequences of not meeting them.

Aspect 4: Test Analysis (MS 13–14)

This aspect focuses on analysing test data to check whether the test is functioning as intended, both overall and for different groups of test takers (see Chapter 6 and Appendix III in this Manual). It involves statistical and qualitative analyses of item and task performance, subgroup fairness checks, and psychometric monitoring. These analyses provide the evidence base for evaluating test quality, identifying problems, and making improvements. Systematic test analysis turns raw performance data into actionable insights, ensuring that the test remains both fair and valid over time.

4.1 MS 13: Test-taker analysis
Standard: You collect and analyse data on an adequate and representative sample of test takers, and can be confident that their achievement is a result of the skills measured in the examination and not influenced by factors like L1, country of origin, gender, age and ethnic origin.

What this means in practice: You must gather demographic and background data and analyse test performance to identify potential bias. Techniques include Differential Item Functioning, subgroup analyses, or multi-group factor analysis (see Appendix III). When statistical approaches are not feasible (e.g. small test populations), qualitative evidence (surveys, interviews, expert reviews) should be used. Where evidence of bias exists, you must act by revising or removing items, adapting specifications, or strengthening training for item writers and raters.

4.2 MS 14: Psychometric analysis
Standard: Item-level and task-level data (e.g. for computing the difficulty, discrimination, reliability and standard errors of measurement of the examination) is collected from an adequate sample of test takers and analysed.

What this means in practice: You must conduct regular psychometric analysis of all test materials using appropriate sample sizes and methods. This includes difficulty, discrimination, reliability, and error of measurement. Classical and modern test theory approaches should be applied where feasible (see Appendix III). Results must inform item bank updates, test specifications, and item writer guidance. If test populations are too small for stable statistics, qualitative alternatives (e.g. expert judgement, small-scale trials) must be employed to monitor quality.

Aspect 5: Communication with Stakeholders (MS 15–18)

This aspect covers how test providers communicate with test takers, institutions, policymakers, and the wider public (see Chapter 2 in this Manual). It includes results reporting, guidance on score interpretation, information about test use, and ongoing stakeholder engagement. Clear and transparent communication ensures that results are not only accurate but also properly understood and responsibly applied. By maintaining trust, accountability, and an open flow of information, this aspect strengthens the social impact and legitimacy of the testing programme.

5.1 MS 15: Results reporting
Standard: The examination administration system communicates the results of the examinations to test takers and to examination centres (e.g. schools) promptly and clearly.

What this means in practice: You must provide results in a secure, timely, and understandable way. Delivery systems (online portals, printed certificates, email) must have safeguards such as access codes or encryption. Score reports must be clear, accurate, and user-tested for comprehensibility. You must meet published timelines for result release, with quality assurance checks to avoid errors. Different audiences (test takers vs. institutions) may require different report formats, so systems must be flexible.

5.2 MS 16: Test information for stakeholders
Standard: You provide information to stakeholders on the appropriate context, purpose and use of the examination, on its content, and on the overall reliability of the results of the examination.

What this means in practice: You must publish accurate and accessible information about your test, tailored to different users. This includes test purpose, content, reliability and validity, and guidance on appropriate uses. Materials should range from plain-language overviews for test takers to technical manuals for expert users. Accessibility requirements (e.g. alternative formats, translations) must be considered. Information must be regularly reviewed and updated.

5.3 MS 17: Score interpretation guidance
Standard: You provide suitable information to stakeholders to help them interpret results and use them appropriately.

What this means in practice: You must explain what scores mean, their limitations, and how they should be used in decision-making. This includes publishing descriptors, scale explanations, error margins, and illustrative examples. Institutions may need predictive validity evidence or guidance for placement. Test takers may need personalised feedback or study recommendations. You must highlight the limitations of scores and encourage multiple sources of evidence for high-stakes decisions.

5.4 MS 18: Preventing test misuse
Standard: You take action when test scores are interpreted and used for a purpose that violates the best interest of test takers.

What this means in practice: You must maintain communication with stakeholders, including test sponsors, organisations using the scores to make decisions, programmes providing test preparation resources, and test takers, to ensure that you are aware of how they are interpreting and using test scores. (For testing programmes with broad uses, monitoring websites and social media concerning

the test may also be helpful.) Be aware of potential misuses and develop action plans appropriate to your testing situation; these may include discussions with score users to clarify appropriate uses, invalidation of scores if the test was given to inappropriate populations, or communications to test takers to inform them of their rights, depending on the nature of the misuse. An example of a misinterpreted test score might be if a test taker has taken a B2 level test in all four skills for a practical job position, yet the test taker is not required to write in their job. The writing part is construct-irrelevant and could negatively affect the interpretation of the test taker's language ability.

Appendix II – Collecting pretest/ trialling information

This appendix contains possible questions to ask of the people involved in pretesting or trialling. (See 3.3.3.)

Note on applicability

The following questions are offered as examples of the type of feedback that may be gathered after pretesting or trialling. They are organised by the groups of people providing the feedback: administrators/proctors, test takers, and examiners or raters. Within those groups, the questions may be divided by their relevance to types of test tasks (for example, oral comprehension or written production). These are only examples: testing programmes should consider what information they need, and may use only some of the questions, or may add others. The questions can be administered as a paper or online survey, or in group feedback sessions or individual interviews, depending on the number of responses that might reasonably be obtained. Ideally, the questions should be posed in the language most familiar to the people providing the feedback; if that is not possible (for example, in an online survey for a test with a population with diverse first languages), make sure the language is simple enough to be understood by all participants.

Pretest/trialling feedback from proctors – all components

Please comment on the test administration process and test-taker experience before, during, and after the test:

1. Was the venue/platform appropriate and user-friendly for delivering the test? For example, was the room quiet and lighting adequate?
2. Did any disturbances occur (noise, interruptions, etc.)? Were there any technical problems with the delivery platform?
3. Had instructions been given to test takers before the test began, as specified? Were the instructions clear?
4. Were there any technical issues (e.g. audio playback, timing problems)?
5. Were test takers notified before the exam regarding adjustment they might need? And were these adjustments provided? If so, what adjustments were provided?
6. Did test takers appear to understand the rubrics and what was expected of them?
7. Was timing appropriate for each component?
8. Were any test takers distressed, confused, or disengaged during the test? If yes, in your opinion what was the reason for this?
9. Do you have any suggestions for improving test-taker experience or logistics?

Pretest/trialling feedback from test takers

A. General

1. Was the test location adequate (room quiet and free of distractions)?
2. (If the test was carried out online) How accessible was the online platform/application?
3. Were the instructions clear?
4. Did you understand what every task was asking you to do?
5. Did you have enough time to complete the task(s)? (If not, how much extra time would you have needed?)
6. Were there any words or expressions in the rubrics or instructions you did not understand which prevented you from completing the task? (If yes, please indicate some examples)
7. Did you feel confident answering the questions? If not, why not?
8. Do you have any suggestions to make the tasks easier to understand or complete?
9. Do you have any additional comments or suggestions?

B. For items/tasks based on written or spoken texts

1. Could you follow the speaker's/writer's ideas and arguments? (Easily-With some difficulty-With great difficulty) For lower levels: Did you understand the general idea of the text?
2. How familiar were you with the topic? (Very familiar-Quite familiar-Not very familiar-Unfamiliar)
3. (For items/tasks based on written texts) Was the way the material was presented easy to read (for example, the way the text and tasks were arranged)?

C. For items/tasks using recorded audio

1. Was the audio clear and at an appropriate volume?
2. Was the speaker's accent understandable?
3. Was the delivery speed appropriate?
4. Did you have enough time to read the questions before and after listening?
5. Were there any non-verbal elements (e.g. laughter, background noise, unnatural pauses) which made you less confident in your responses?

D. For tasks involving written or spoken responses

1. Was the topic clear and familiar to you?
2. Did the topic feel relevant to your experience?
3. Was the task explained clearly?
4. Were the visuals used helpful (e.g. pictures, questions, text)?
5. Were you sure about:
   a. the format of your writing (e.g. letter, essay)?
   b. the appropriate level of formality?
   c. the target reader/listener?
6. the length required (number of words to write/how long to speak)?

7. Did you have enough time to plan and write your answer/to prepare what to say?
8. Did you feel you had enough time to express your ideas?

E. For tests with interlocutors (examiners and/or interlocutor test takers)

1. Was the attitude of the examiner professional and supportive?
2. Do you feel the examiner was directing your responses appropriately? (for example, not interrupting, not filling things in for you, asking follow-up questions that make sense)
3. Were the other speaker(s)' accents understandable?
4. If the task was interactive, were you able to understand and interact with the other speaker(s) easily? If not, why not?

Pretest/trialling feedback from examiners or raters – written or spoken responses

About the task **input**: Focus

1. Was the task generally understood?
2. Was the test taker's role as a writer or speaker (e.g. someone providing information, presenting a case) clearly identified?
3. Was the target reader/target audience clearly identified?
4. Was the register appropriate for the task?
5. Does the task elicit enough language to be rated? At the intended level? Does the task elicit the communicative/pragmatic functions intended?
6. Did the task allow variation in discourse structure, vocabulary, and fluency?
7. Is there any evidence of cultural bias? Does the task favour a test taker of a particular background or age? Could it be traumatising or unsettling for some test takers?
8. Was the number of tasks adequate?
9. Were the tasks sufficiently varied? (e.g. in topics, expected type of response)

About the task **input**: Language

1. Was the question/task understandable by the target level test takers?
2. Was there any confusion over or misinterpretation of the wording?
3. Were test takers confused about the appropriate register to use?
4. Was the question free from ambiguity or misinterpretation?
5. Should the writing task instructions or prompts be reworded? If so, please outline your suggestions.

About the **interaction** with the examiner, if applicable

1. Was the examiner's behaviour appropriate? (not giving up too easily, not prompting too much, not speaking too much, giving feedback)
2. Did the materials provide enough guidance for structuring the interaction with the test taker(s)?

About the task **output**: Content

1. Was the task type interpreted appropriately?
2. Were any content points misinterpreted/omitted? Please provide examples.
3. Was the word length/speaking time limit suitable for the task?

About the task **output**: Range/tone

1. Was any language lifted from the question? Please specify.
2. Did the test takers use the communicative and pragmatic functions as intended?

About the task **output**: Level

1. Did the task allow test takers at the targeted level to demonstrate their competencies?
2. Was the task format effective in eliciting written/spoken performance at the intended CEFR level?
3. Rating scale:

   Please outline suggestions for improving the clarity or application of the rating scale.
4. Overall impression:

   Please summarise your evaluation of this writing/speaking task.

Do you have any suggestions for:

- improving prompt design
- adapting format for greater authenticity
- improving rating procedures?

Appendix III – Using statistics in the Testing Cycle

Introduction

Using statistics in the Testing Cycle is an important way to gather evidence regarding the validity, the reliability and the fairness of language exams. This appendix describes some statistics that can be used with language testing data. It contains five sections, each covering a different topic. Section 1 considers the testing population, which is important for any further statistical work; Sections 2 and 4 cover different techniques for determining the measurement characteristics of tasks and items; Section 3 looks at reliability, both for the test as a whole and for subjectively rated items; and Section 5 introduces a technique (Differential Item Functioning) for assessing bias. For additional topics, such as factor analysis of test tasks, and a more in-depth treatment of statistical analyses written for non-psychometricians, please refer to Bachman (2004). For statistical analyses specifically related to standard setting, see the *Manual for Relating Language Examinations to the CEFR* (Council of Europe, 2009), Cizek (2001), and Cizek and Bunch (2007). Note that some statistics can be applied with good success by non-statisticians, whereas others may need more technical support. This Appendix is not an exhaustive guide to how to run the analyses; rather, it provides context and guidelines to help practitioners determine which analyses may be needed and to interpret analyses appropriately.

Before diving into specific statistical procedures used in language test analysis, it is important to offer a few guiding principles, so results can be interpreted appropriately and to avoid common misunderstandings. Statistics are powerful tools, but they must be used thoughtfully and interpreted in the context of the exam.

Descriptive vs. inferential statistics

Descriptive statistics summarise the characteristics of a dataset. They include measures such as averages, percentages, ranges of scores (minimum and maximum), and standard deviations (SD, measuring how much scores are spread out from the average), and are used to describe what has been observed, for example, the average score on a test or the distribution of scores across test takers. Descriptive statistics focus on the population of the dataset. One can check, for instance, if the mean score of a subgroup of the test takers who took the test (e.g. women) is higher or lower than the mean score of another group (men). At this stage, it is simply an observation based on the data.

Inferential statistics, on the other hand, allow users to make generalisations or predictions about a larger population based on a sample. These include hypothesis tests, confidence intervals, and regression models. They rely on probability and are subject to uncertainty. For instance, inferential statistics make it possible to determine (with some degree of uncertainty), whether the observed mean difference between two subgroups is likely to reflect a real difference in the population, and not just a random fluctuation in the sample. A statistical test (in this case a t-test might be performed), allows us to infer whether the difference is statistically significant.

But such a statistical test (like many statistical methods or models) relies on assumptions about the data, such as the normality of the distribution in the population and the independence of the observations. If these assumptions are violated, results may be misleading. The interpretation of the results must also take into account the context in which the sample data was collected. In the previous example, if the statistical test is significant, claiming that women score differently on the test than men assumes that the test takers who took the test can be considered as a random selection of the target population. If all women were randomly picked from a specific institution and men randomly picked from another institution, one can only claim that women from the one institution score differently than men from the other: the differences might be due to curricula at the different institutions, or some other factor, rather than gender alone.

Sample size, significance, and effect size

A statistical test compares a value calculated from a dataset to the range of values that could be observed from random samples of the same size drawn from a larger population. The more the value calculated from the dataset differs from the range of values expected from random samples, the more likely the observation isn't due to chance. Statistical significance (often expressed through *p*-values) tells whether an observed effect is likely due to chance.

However, significance alone does not indicate the importance of an effect. Sample size plays a major role: with large samples, even very small effects, which have no practical consequences, can appear statistically significant. Effect size quantifies the magnitude of a result and helps assess its practical relevance. On the other hand, notably with small samples, an effect can be important enough to merit attention (and prompt further investigations), even while not being statistically significant. Therefore, both significance and effect size should be considered together, especially in the context of test design and decision-making. Note also that what constitutes an adequate sample size differs depending on the particular analysis being performed. In particular, Item Response Theory (IRT) models (see Section 4) typically require much larger sample sizes to be useful than classical item analysis (see Section 2). In determining which analyses to use, testing professionals must balance the type of information needed, sample size limitations, and strengths and weaknesses of particular methodologies.

Assumptions and model fit

As said earlier, many statistical methods rely on assumptions about the data, which must be checked to correctly interpret the results. When using statistical models (e.g. regression, Item Response Theory), it is also essential to check whether the data fits the model. Using a model which is not in line with the data can lead to invalid conclusions. Determining model fit is highly technical, but graphical observations, goodness-of-fit indicators and diagnostic checks help ensure that the conclusions drawn are valid. See Bachman (2004, pp.283–285) for an example of trying different models to fit the data. Note that once overall model fit is established for the dataset, it is common for some items or raters to be 'outliers' or 'misfitting,' and in these cases the lack of fit to the model can indicate a potential problem with the item or rater (see Section 4).

Choosing the right statistic and interpreting it in context

There is rarely a single 'correct' statistic for a given situation. Different methods may highlight different aspects of the data or even provide contradictory results. Some methods have been used for decades and are widely recognised, whereas others may offer new insights, but their value may not be clear. Therefore, it is recommended to use different statistics for a given situation to be more confident about the claims that can be made. Moreover, rules of thumb (e.g. a reliability coefficient above 0.80) should not be applied mechanically. The context of the exam — its stakes, purpose, and population — must guide interpretation. The purpose of the statistical analysis is also important: for routine review and potential revision of individual test items, it might be useful to see difficulty and discrimination analyses for incorrect responses (see Chapter 6 on review and revision), whereas for decisions about whether an item should count toward a score, the overall difficulty and discrimination is likely adequate. For these reasons, it is recommended, when investigating a specific point, to make a review of the commonly used indexes in the community of language testing and the different reference values considered, and to use critical thinking to select the ones that will be most useful, keeping in mind the specific test context. Many resources exist to explain statistical results in plain language, including tools within some testing platforms, or asking AI assistants for help.

References

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge University Press.

Cizek, G. J. (Ed.). (2001). *Setting performance standards: Concepts, methods, and perspectives*. Erlbaum.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Sage.

Council of Europe. (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR) – A manual*. Available online: https://www.coe.int/en/web/common-european-framework-reference-languages/relating-examinations-to-the-cefr

Section 1: Describing the test population and checking for representativeness

When any kind of analysis or research is done using test data, the data should 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.

Collecting and analysing test data requires planning and resources, 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).

Data on the background characteristics of test takers can be collected whenever a test version is administered (see Sections 3.3.3 and 4.2.1). Characteristics can be compared using simple percentages, e.g. comparing the balance of males with females in two different samples. Inferential statistics help establish whether any differences between two samples are likely to be due to chance.

In this section, we will see how to use statistics to describe the population of a test, based on some demographic and score information, and to assess the representativeness of a sample under study with regard to this population.

Describing the test population

The first step in analysing test results is to describe the population of test takers who took the test during a sufficient time period, where no changes in the population have been noticed, to have a comfortable amount of data. This provides essential context for interpreting results and for comparing with other groups, such as an experimental or a pretest sample. As an example, assume that there is at hand a test population consisting of 1,000 test takers, with some background data. We can summarise key characteristics as follows:

Variable	Example statistics
Gender	600 female (60%), 400 male (40%)
Age	Mean = 23.5 years, SD = 4.2 years, Range = 18–40
Country	France: 500 (50%), Belgium: 300 (30%), Others: 200 (20%)
Test Score	Mean = 72, SD = 8, Median = 73, Range = 50–90

These descriptive statistics provide a snapshot of the test takers. For example, we see that most test takers are female, the average age is about 23, half are from France, and the average test score is 72. This is just an observation based on the data. No inference is made about other populations. Graphs such as bar charts (for gender and country) and histograms (for age and scores) can help make these distributions clear and accessible.

Checking representativeness

Suppose a new sample of 150 test takers is used to conduct research or to pretest new items. To ensure that the results of the research can be generalised to the test population, it is important to check whether this group is representative of the main test population. To do so, the same descriptive statistics can be calculated and compared to the ones of the test population:

Variable	Test population	Research sample
Gender (% female)	60%	58%
Mean age	23.5	23.1
% France	50%	52%
Mean score	72	71

At first glance, the sample appears similar to the main population. Inferential statistics can then be used to test if differences can be considered meaningful, or just due to chance. In this example, the following tests can be used:

a) Comparing proportions (e.g. gender, country)

- Test: Chi-square test of independence
- Null hypothesis: The proportion of females (or test takers from France) is the same in both groups

b) Comparing means (e.g. age, test score)

- Test: Independent samples t-test
- Null hypothesis: The mean age (or mean score) is the same in both groups

The 'null hypothesis' is a statement that the differences are just due to chance. The tests produce *p*-values, which show the probability of obtaining the data if the null hypothesis were true. A very low *p*-value (below 0.05) would show that it is *unlikely* that the results are due to chance. Suppose the tests yield the following *p*-values:

Variable	Test used	p-value	Interpretation
Gender	Chi-square	0.68	No significant difference
Age	t-test	0.42	No significant difference
Country	Chi-square	0.55	No significant difference
Test score	t-test	0.09	No significant difference

All *p*-values are above 0.05, so we do not reject the null hypothesis for any variable, meaning that the differences between populations are likely due to chance. This suggests that the research sample is statistically similar to the test population in terms of gender, age, country, and test scores. Note that, in addition to *p*-values, it is good practice to report effect sizes (e.g. Cohen's *d* for means, Cramér's V for proportions) to assess whether any differences, even if not statistically significant, might be practically important, notably if the sample is small.

Free statistical tools to run such analyses

Stats.Blue (https://stats.blue/)

Statisty (https://statisty.app/), can be used online or downloaded

JASP (https://jasp-stats.org/), more complete, to be downloaded

References

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge University Press.

Green, R. (2013). *Statistical analyses for language testers*. Palgrave Macmillan. https://link.springer.com/book/10.1057/9781137018298 [for users of SPSS software]

Schneider, G., & Lauber, M. (2020). *Statistics for linguists: A patient, slow-paced introduction to statistics and to the programming language R*. University of Zurich. https://dlf.uzh.ch/openbooks/statisticsforlinguists/ [FREE, for people who want to practise with R]

Section 2: Classical item analysis

More precise data on test taker performance, notably item response data (the response made by a test taker to each item in the test), can show how well items worked and where the focus of editing should lie. The two primary methods for analysing this type of data are classical item analysis (CTT) and Item Response Theory (IRT; see Section 4). CTT focuses on the total test score and assumes that each test taker's observed score is composed of a true score and random error. It is widely used due to its simplicity and ease of application. It works better in the case where items are parallel (i.e. have similar measurement properties, such as mean and variance) and can be considered interchangeable, which is more the case for exams targeting a specific level than for a multi-level test.

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, but with more uncertainty about claims that can be made – see the section about sample size at the end of this section).

Classical item analysis is used:

- To analyse pretest data (see Section 3.3.3) and inform the selection and editing of tasks for live use
- To analyse live test response data (see Section 6.1)

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
- (For multiple-choice items) 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 acceptable values for some of these statistics are. They are not intended as absolute rules – in practice the values typically observed depend on the context. Classical statistics tend to look better when there are:

- Larger numbers of items in a test
- More test takers taking the test
- Wider ranges of actual ability in the group taking the test

Conversely, they look worse when there are few items or test takers, or a narrow range of ability. So it is important to always keep in mind the context of the test when interpreting the results.

Several free software packages can be used to run classical analysis (see below). The table below shows example item statistics based on output from the ShinyItemAnalysis software. It shows the analysis for three items.

	Avg. score	SD	RIT	RIR	ULI
Item 1	0.525	0.499	0.379	0.230	0.427
Item 2	0.570	0.495	0.367	0.218	0.424
Item 3	0.714	0.452	0.272	0.131	0.291

Facility

Facility is the proportion of correct responses (Avg. score in the table). 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. The table 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.

In the case of an exam targeting one specific level, if the group of test takers 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.

In the case of a multi-level test (for instance a B1 to C1 test) where the test population comprises many advanced learners, it is common to have some items with a facility greater than 0.80, because an item targeting the low end of the range will generally be answered correctly by all test takers except those at the low end. These items are nevertheless necessary to cover the construct of the test and to classify low-level test takers. Context is key to correctly interpreting data.

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 (in the table above, the discrimination index is labeled Upper-Lower Index (ULI), and there are two different point biserial calculations, labeled RIT and RIR).

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). But discarding a third of the data may be problematic if the sample of test takers is already small. This index is not always available in software but can easily be calculated by hand.

The output in the table above shows these in the ULI column. For Item 1 the value of the discrimination index is 0.427.

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, both the low and high groups will score well (or badly), and thus discrimination will be underestimated by this index.

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 either the whole test – this is RIT – or on the rest of the items of the test, that is, all items except for the one in question – this is RIR. It is very often calculated by software dealing with item analysis. RIR is generally preferable, as it avoids having the item in question inflate the correlation.

In general, a point biserial correlation of greater than 0.30 is considered strong and a correlation between 0.20 and 0.30 is considered moderate. Point biserials between 0 and 0.20 are weak, though the item is still providing some information. A point biserial below 0 means that strong test takers

are more likely than weak test takers to get the item wrong, so the item is actively damaging the measurement. In such a case, check whether one of the distractors is actually a correct answer, or the key is wrong. Even if the item appears correct, however, if the point biserial is negative the item should not be used.

The output in the table above shows a low RIR for Item 3 (0.131), which should be investigated. The fact that RIT values are clearly higher than RIR values in this example is due to the limited number of items in the dataset (20), as the effect of including the item in question is magnified the fewer items there are in the dataset.

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. A distractor analysis shows the proportion of test takers choosing each distractor.

Below is another output table from the ShinyItemAnalysis software, which shows the % or endorsement of each option (A, B, C, or D) of Item 1 in the dataset among all the test takers.

Response	Group 1
A	0.53
B	0.21
C	0.18
D	0.08

Option A is the correct option for Item 1, and its facility is 0.53. Distractors B and C attract test takers to a similar degree, as about the same proportion of test takers chose B chose C. Distractor D is clearly less attractive and would deserve editing if the item was much easier than expected. 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.

Graphical displays such as Figure 14 are generally useful to run distractor analysis and help diagnose potential problematic items. In this figure, the test takers are divided into four groups of comparable size by their total score (the number of groups can be adjusted to the size of the sample). The curves show the percentage of endorsement of each option by each group or test takers. For a good item, the correct-answer curve should be increasing from Group 1 to Group 4 (and usually the steeper the increase, the more discriminating the item) and distractor curves should be decreasing.

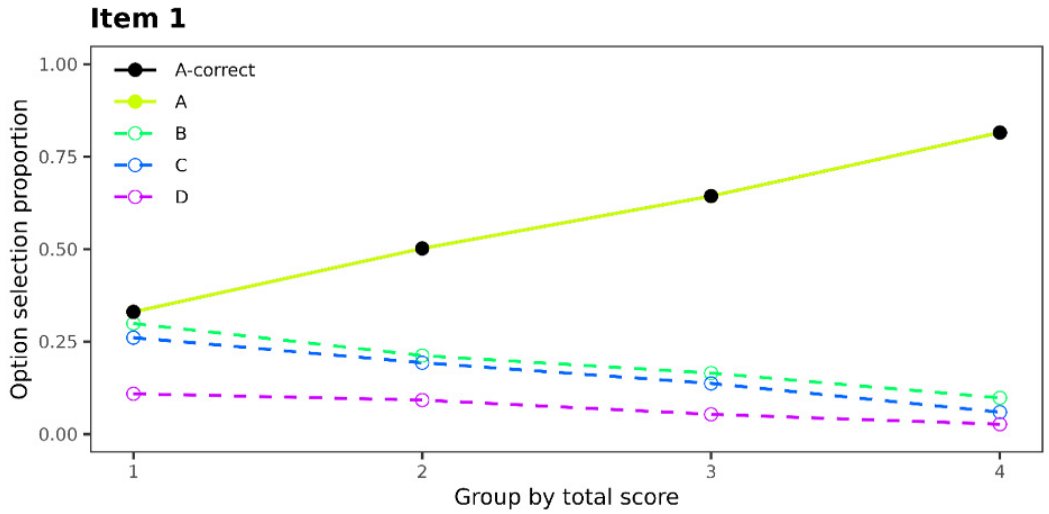


Figure 14 ***Example item graphical analysis (ShinyItemAnalysis)***

How to interpret the results given the sample size

Reference values are helpful when stating the quality of an item, but they should be interpreted keeping in mind the context of the exam, and notably the sample size.

Figure 15 shows, for different sample sizes (on the x axis), the lower and upper value (on the y axis) of a 95% confidence interval for an item whose facility value is estimated to 0.81 (following Verhelst, 2004a). When the sample size is 200, one can be confident that the facility of the item for the population is between 0.75 and 0.86, so one may decide to reject this item because it seems too easy for test takers. But when the sample size is 20, the lower bound of the confidence interval is 0.59. As a consequence, there is a higher risk of rejecting an item of acceptable facility just because of a sampling effect.

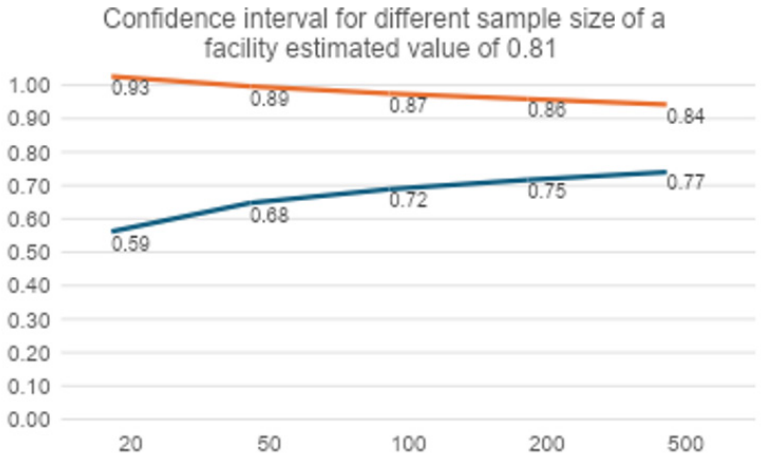


Figure 15 ***Impact of sample size on decisions regarding item facility***

Figure 16 shows, for different sample sizes (on the x axis), a 95% confidence interval (on the y axis) for an item whose point biserial (Rpbis) value is estimated to 0.3. It appears clearly that if the sample size is 500, the point biserial for the population will lie most probably between 0.2 and 0.4, showing an acceptable correlation between the item score and the total score. But when using such a reference value for a sample size of 50, one takes a bigger risk of selecting an item with a weak correlation with the total score, possibly even a negative correlation.

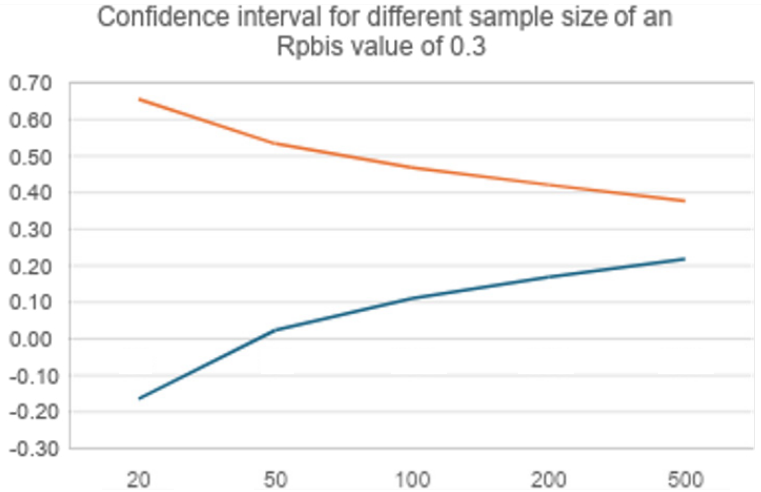


Figure 16 ***Impact of sample size on decisions regarding item discrimination***

Free statistical tools to run classical item analysis (and more):

- CITAS (https://assess.com/citas/) [using Excel, limited to 100 items and 100 test takers]
- Iteman (https://assess.com/iteman/) [downloadable software, limited to 100 items and 100 test takers]
- ShinyItemAnalysis (https://shinyitemanalysis.org/) [can be used online or as a R package, with no limitations]
- jMetrik (https://itemanalysis.com) [downloadable software based on Java, more comprehensive and more complex]

Reference

McCowan, R. J., & McCowan, S. C. (1999). *Item analysis for criterion-referenced tests*. Available online: https://files.eric.ed.gov/fulltext/ED501716.pdf

Verhelst, N. (2004a). *Reference supplement to the preliminary pilot version of the manual for relating language examinations to the Common European Framework of Reference for Languages: learning teaching and assessment. Section C: Classical Test Theory*. Available online: https://rm.coe.int/1680459faa

Section 3: Estimating reliability of test scores (items and task rating)

Reliability is an essential characteristic for the valid use of test scores: the meaning of a score or grade is unclear if items can be answered correctly by random chance, or raters rate inconsistently. There are many different statistical approaches to reliability, and different statistical analyses are used for estimating reliability in different contexts. This section discusses overall test reliability (see Section 1.5.2), typically measured in terms of the consistency of performance of test takers on components of the test, and the reliability of rater judgements (see Section 5.2), measured in terms of agreement across different raters and the internal consistency of individual raters' judgements.

Overall test reliability

There are several ways in which to estimate reliability of test scores and several formulae to do so. Each method has different assumptions, and the relevance of using some widely used indexes, such as the well-known Cronbach alpha, may be questioned (Sijtsma, 2008; Béland et al., 2017) in some contexts. Therefore, it is important to make a review of existing indexes to identify the ones which are relevant for the context of the exam (see, for example, Béland & Falk (2022)). It is also good practice to make use of several indexes to get a more comprehensive view of the behaviour of the exam.

As an illustration, we can cite the split half method, which 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 that the two halves be as equivalent as possible: they cover equivalent areas of the construct, they are of equivalent difficulty, etc. The more the scores on both parts are similar, the greater the estimate of their reliability.

Reliability and standard error of measurement

Once the reliability of test scores has been estimated, it is important to assess the adequacy of the value **in the context of the exam**. Indeed, reliability estimates are dependent on the population of test takers (Verhelst, 2004a). One formula for the definition of reliability is:

$$\text{Reliability} = 1 - \frac{\text{Var}\,(E)}{\text{Var}\,(X)}$$

Where Var(E) is the error variance and Var(X) is the variance of observed scores.

This formula shows that, for the same test, a larger variance on test scores will lead to a higher estimate of reliability. This is why it is important to think critically when using reference values. A test which is designed to classify test takers from B1 to C2 levels will clearly have a better reliability estimate if the population is heterogeneous (in terms of language ability) than if most of the test takers are C1 or C2.

One way to evaluate the reliability measure with respect to the population is to calculate the Standard Error of Measurement (SEM) and to gauge whether such an amount of error is acceptable for the decisions one wants to make.

The SEM is given by the formula:

- SEM = sqrt(Var(X)*(1-reliability))

It gives an idea of the measurement error 'on the average' for test takers and allows us to build confidence intervals around test scores. One can see that, for two test samples having the same variance of observed score, the one with the higher reliability will have a smaller standard error of measurement. But conversely, for two test samples having the same reliability estimate, the one with the more homogeneous test scores will have a smaller standard error of measurement.

One inconvenience of the SEM is that it suggests that the amount of error is the same all along the measurement scale, which is generally not the case. For non-binary classification of test takers (for example, for multilevel tests), one might be more interested in knowing the amount of error at each cut-off point. Models of measurement of the family of Item Response Theory (IRT), like the Rasch model, give the possibility to estimate such Conditional SEM.

Reliability and rater agreement

The estimation of reliability of test scores for rated components of a test is typically calculated using different tools from those used to calculate score reliability for objectively marked components. As explained in Section 5.2, different aspects of the testing situation impact the reliability of such tests, which cannot be reduced to the agreement between raters.

Rater agreement is nonetheless a cornerstone of reliability. If a group of raters score a common sample of oral or written productions very differently, it will be difficult to trust in the scores delivered to the test.

The performance of raters can be assessed statistically in a very simple way by calculating the mean and standard deviation of their ratings (recall that standard deviation is a measure of the spread of their ratings). 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 (notably when the range of scores is large), any number lower than 0.8 should be investigated further, as it suggests that

the raters did not perform similarly. But this correlation can be high even when the two raters show different severity trends, so a number higher than 0.8 does not always signal good agreement. Note also that some variation is expected between humans, and very high values, above 0.95, may indicate that the raters are not independent.

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 (see Section 4). 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. (Each rater has a 'measure' in logits, so that raters with high negative measures are very lenient, and raters with high positive measures are very severe.)

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 separate pockets of data are formed, such as separate pools of raters who do not rate any tasks in common, MFRM may not be able to provide estimates for all elements.

It is important to remember that, in order to be meaningful, MFRM analysis requires independent ratings. This means that it cannot be used when ratings are arrived at through consensus, or where the second rater can see the ratings that the first rater awarded, as knowing the previous ratings might influence the second rating.

MFRM provides information about the relative severity of raters that is very useful for feedback and training. It can also be used to adjust scores taking rater severity into account. With regard to using these scores, known as 'fair average,' it should be noted that MFRM, like any Rasch analysis, assumes that the observations are 'locally independent' from one another. This is a statistical requirement which is difficult to meet and which, when violated, will lead to an overestimation of the reliability of the ratings. Readers interested in following up on this are referred to Verhelst (2004b) and Eckes (2009).

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)

When evaluating the reliability of automated rating tools, additional analyses such as quadratic weighted kappa may be needed (see Williamson et al., 2012).

References

Béland, S., & Falk, C. F. (2022). A comparison of modern and popular approaches to calculating reliability for dichotomously scored items. *Applied Psychological Measurement*, *46*(4), 321–337. https://doi.org/10.1177/01466216221084210

Béland, S., Cousineau, D. & Loye, N. (2017). Utiliser le coefficient omega de McDonald à la place de l'alpha de Cronbach. *McGill Journal of Education / Revue des sciences de l'éducation de McGill*, *52*(3), 791–804. https://doi.org/10.7202/1050915ar

Eckes, T. (2009). *Reference supplement to the Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: learning teaching and assessment. Section H: Many-Facet Rasch Measurement*. Available online: https://rm.coe.int/1680459faa

Linacre, J. M. (2009). *Facets 3.64.0* [Software]. https://www.winsteps.com/index.htm

Sijtsma K. (2008). On the use, the misuse, and the very limited usefulness of Cronbach's Alpha. *Psychometrika, 74*(1), 107–120. doi: 10.1007/s11336-008-9101-0

Verhelst, N. (2004a). *Reference supplement to the preliminary pilot version of the manual for relating language examinations to the Common European Framework of Reference for Languages: learning*

teaching and assessment. Section C: Classical Test Theory. Available online: https://rm.coe.int/1680459faa

Verhelst, N. (2004d). *Reference supplement to the preliminary pilot version of the manual for relating language examinations to the Common European Framework of Reference for Languages: learning teaching and assessment. Section G: Item Response Theory*. Available online: https://rm.coe.int/1680459faa

Williamson, D.M., Xi, X., & Breyer, F.J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2–13.

Section 4: IRT modelling and Rasch analysis

As noted in Section 2, there are two main measurement theories used in the field of language testing: Classical Test Theory (CTT; see Section 2) and Item Response Theory (IRT). Recall that CTT assumes that each test taker's observed score is composed of a true score and random error, and that it works better when items have similar measurement properties and can be considered interchangeable, which is more likely for exams targeting a specific level than for a multi-level test.

IRT can provide a better understanding of item difficulty (see Section 3.4.2) than classical analysis and has more additional applications, such as linking test forms; it is also more suitable than CTT for use with multi-level tests. The application of IRT, however, requires the use of specific software and makes stricter assumptions on the data. IRT models the probability of a test taker answering an item correctly based on the test taker's ability level *and* the properties of the item (difficulty, sometimes discrimination). The simplest and most used IRT model in language testing is the Rasch model. In the Rasch model framework, items are characterised by a difficulty parameter. Other IRT models use more parameters to characterise items and therefore need a bigger sample to provide consistent estimates. Therefore, when the assumptions of the Rasch model are met (notably the fact that items have a comparable discrimination power) and when data fits such a model reasonably well, it should be preferred to other IRT models.

With Rasch analysis:

- The precise difference in difficulty between two items is clear, as items are placed on an interval scale measured in logits (a scale with a mean of zero and values generally between –3 and +3): this measurement can be used across different administrations of the test or test version
- 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 or test versions must be linked. For example, two tests or test versions may be linked in a number of ways:

- Some common items are used in both tests or test versions
- A group of anchor items are administered with both tests or test versions
- Some or all items are calibrated before being used in a live test (see Section 3.3.3)
- Some test takers may take both tests or test versions

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

Standards can be monitored (see Section 6.1) 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 or test versions are constructed using calibrated items. The overall difficulty of a test or test version can be described with its mean difficulty and the range. The difficulty can be controlled by selecting a group of items that fit the target range and match the target mean, or by producing a Test Information Function (TIF), which maps the distribution of measurement across the test's difficulty range.

When items are first calibrated, the difficulty values will not mean much. But over time it is possible to study the real-world ability of test takers in order to give meanings to points on the ability scale. Alternatively, or at the same time, subjective judgements about the items ('I think a threshold B1 learner would have a 60% chance of getting this item right') can be used in order to give meaning to the item difficulty values. In this way the numbers on the ability scale become familiar and meaningful.

Minimum number of test takers for Rasch analysis

For probabilistic models, it is generally true that the accuracy of the calculations is significantly affected by the size of the population. This is so true that one of the drawbacks of IRT models is precisely the need for large samples (see e.g. Coulacoglou & Saklofske, 2017). For IRT models, there is no consensus in the literature on the minimum acceptable population size, except that there is agreement that a one-parameter model can be used for smaller populations than a two- or three-parameter model. Wright and Stone (1979) suggest a lower limit of 200, while McNamara (1996) suggests that 100 is sufficient, although he emphasises that larger populations increase the precision of the analysis. However, according to Embretson and Reise (2000), the robustness of parameter estimation is not guaranteed even for 350 individuals.

Free software for IRT modeling

- ShinyItemAnalysis
- jMetrik
- R, with Dexter (and its graphical interface dextergui) library, or eRm for Rasch modelling, mirt for other IRT models

References

Andrich, D., & Marais, I. (2019). *A Course in Rasch Measurement Theory: Measuring in the Educational, Social and Health Sciences*. Springer Link.

Coulacoglou, C., & Saklofske, D. H. (2017). *Psychometrics and Psychological Assessment: Principles and Applications*. Academic Press.

Dionne, E. & Béland, S. (2023). *Appliquer le modèle de Rasch: Défis et pistes de solution*. Presses de l'Université du Québec.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Erlbaum.

McNamara, T. (1996). *Measuring second language performance*. Addison Wesley Longman Ltd.

Verhelst, N. (2004d). *Reference supplement to the preliminary pilot version of the manual for relating language examinations to the Common European Framework of Reference for Languages: learning teaching and assessment. Section G: Item Response Theory*. Available online: https://rm.coe.int/1680459faa

Wright, B. D., & Stone, M. H. (1979). *Best Test Design: Rasch Measurement*. Mesa Press.

Section 5: Checking for bias with analysis of Differential Item Functioning (DIF)

One element of fairness in testing is detecting whether some groups of test takers have an advantage over others that is unrelated to the construct (see Section 1.5.1.3). After a test has been administered, an effective method for detecting and potentially reducing bias is to examine test items for Differential Item Functioning (DIF). DIF refers to the phenomenon where different groups of test takers with the same underlying ability level perform differently on specific test items. This can indicate potential bias in the test items, which may unfairly advantage or disadvantage certain groups based on characteristics such as gender, age, the format used for the test (digital vs paper-based), or any other construct-irrelevant characteristic. However, note that DIF requires relatively large group sizes for each characteristic, so that in practice, depending on the population, it is not always possible to investigate DIF for all characteristics of interest.

DIF methods help identify items for which individuals from different groups (e.g. males vs. females) with the same ability level have different probabilities of answering correctly. The two primary methods for calculating DIF are the Mantel-Haenszel (MH) non-parametric method and Item Response Theory. When sample size is sufficient and conditions for a Rasch modelling are met, there are sound reasons for favouring the Rasch model over the MH method to detect uniform DIF (Linacre & Wright, 1987), although MH is still used in large scale contexts (Zwick, 2012). For detecting (more complex) non-uniform DIF, one can turn to logistic regression models (Zumbo, 1999) or, when sufficient data is available, to IRT models with two or three parameters when appropriate.

Whatever the method used (Magis et al., 2010), several steps should guide the analysis:

1. **Identify** the groups to be compared (e.g. males and females) and organise the data so that this information is accessible for the analysis.
2. **Conduct a preliminary analysis of the data**. This analysis is a 'trial run' to ensure that the model can be used for the particular dataset. After running this preliminary analysis, check that the dataset is consistent and deal with missing data and outliers. For misfitting items, check to ensure that data was correctly entered (for example, that the item key is correct) and discard any potentially defective items (for example, two plausible answers for a multiple-choice item). Discard items from the DIF analysis that could distort the results (for example, items that are too easy or too difficult may not provide useful information or may skew the results and mask the performance of more problematic items). Also check that the groups being compared are of sufficient size (see the sections on sample size in the Introduction and Sections 2 and 4) to allow for meaningful statistical comparisons and choose appropriate methods (small group sizes can lead to unreliable DIF detection for some methods). It is also important to ensure that the assumptions of the measurement model are met (e.g. unidimensionality).
3. **Apply one or more DIF detection methods and identify potentially problematic items**. Different methods will give partially different results, so it may be useful to cross-reference the results before taking a decision on an item (if an item is identified by all the methods as potentially problematic, that is a stronger indication that it should be revised than if only one method indicates that the item functions differently between the groups).
4. **Review and revise items**. Examine flagged items for content bias and take a decision concerning these items (revise or remove biased items). As there may be statistical interaction between the items, a common approach is to review items with the largest DIF first, remove any of those that need to be discarded, and then reanalyse the resulting new dataset.

While DIF analyses are useful for identifying potentially biased items, the mere presence of DIF does not automatically indicate bias. Items flagged for DIF require further examination by expert panels to determine the underlying features that may have caused the DIF. It is also important to note that not all group differences signify unfair bias. For instance, learners with different first languages (L1s) may find certain items more challenging due to inherent differences between their native languages and the target language. In the context of assessing language proficiency, such discrepancies are

regarded as a natural aspect of measuring proficiency in the target language rather than flaws in the test itself.

Free software to run DIF analyses

- ShinyItemAnalysis (various methods)
- jMetrik (MH method)
- R, with difR library (various methods)

References

Ferne, T. & Rupp, A. A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly*, *4*(2), 113–148. doi: 10.1080/15434300701375923

Linacre, J. M. & Wright, B. D. (1987). *Item bias: Mantel Haenszel and the Rasch Model.* https://www.rasch.org/memo39.pdf

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods 42*, 847–862. doi: 10.3758/BRM.42.3.847

Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores*. Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zwick, R. (2012). *A Review of ETS Differential Item Functioning Assessment Procedures: Flagging Rules, Minimum Sample Size Requirements, and Criterion Refinement*. RR-12-08. ETS. Available online: https://files.eric.ed.gov/fulltext/EJ1109842.pdf

Appendix IV – 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.

adaptive testing
In an adaptive test, after a routing item/set of items, the test taker is presented with an item/a set of items targeting their expected ability based on their previous answers. In a pure adaptive test, ability is estimated after each answer, whereas in a multistage adaptative test, ability is estimated after the test taker has answered all the items in the current stage. This type of adaptivity is score-based, and it is the most common type. Also known as *Computer Adaptive Testing (CAT)*. If the calculation is done with Item Response Theory (IRT), it is also referred to as IRT-adaptivity.

(CEFR) alignment
A process to provide evidence that a test's scores can be reliably and meaningfully interpreted using proficiency descriptors or some other external set of categories, for example, the CEFR or government-established proficiency categories.

analytic scale
An instrument of assessment used for rating written and spoken productions consisting of bands to describe each level of performance, broken down in specific aspects of language (criteria), e.g. for writing: task accomplishment, text organisation, coherence, accuracy, etc. The rater gives marks according to each of the aspects included in the assessment instrument, with the possibility of aggregating them subsequently into a general mark. Sometimes referred to as grid. Compare: **holistic scale**

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
In language testing, the measurement of one or more aspects of language proficiency by means of some form of test or procedure.

authenticity
The degree of correspondence of the characteristics of a given language test task to the features of a target language use (TLU) task (Bachman and Palmer, 1996, p. 23). For example, on a test of listening comprehension for academic purposes, a task in which test takers listen to a lecture and take notes is more authentic than one in which they answer multiple-choice questions. Authenticity is a continuum, with tasks being more or less similar to those in TLU, in terms of context and/or the needs and interest of the language user. 'Genuineness is a characteristic of the passage [the text] itself and is an absolute quality. Authenticity is a characteristic of the relationship between the passage and the reader and it has to do with appropriate response' (Widdowson, 1978, p. 80). A text can be genuine without being authentic in relation to a specific context and/or the interest or needs of the language user. At the same time, a text can be authentic without being genuine if it was edited for adaptation to the task. Compare: **genuineness**

automarker
A computer application, often powered by artificial intelligence, that provides ratings for constructed responses.

background characteristics
Demographic information (e.g. age, gender, L1, country of origin, education, professional activities, etc.) of the population for which the test is produced. This type of data might also be collected (e.g. during registration, through post-test questionnaires) to be used in analysis to demonstrate that the population taking the test is the same as the intended one. Also known as *test taker characteristics*.

band
In its broadest sense, part of a scale. In an item-based test this covers a range of scores which may be reported as a grade or a band score. In a rating scale designed to assess a specific trait or ability such as speaking or writing, a band normally represents a particular level.

bias
A test or item can be considered to be biased if one particular section of the test taker

population is advantaged or disadvantaged by some feature of the test or item which is not relevant to what is being measured. Sources of bias may be connected with gender, age, culture, schooling, etc.

branching adaptivity
Adaptivity where test takers choose their own path through the test (e.g. a test taker is offered three texts and can choose which of the three texts they want to answer questions about). In branching adaptivity, the path is not influenced by the score(s) of the test taker.

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

calibration
The process of determining the positioning of a test on an existing scale (so that different versions of the test may be meaningfully compared, or, if the scale covers more than one level, the distance in terms of difficulty of two versions of different-level tests can be assessed).

clerical marking
A method of marking in which human markers or automated systems grant points mainly for objective items. They mark by following a mark scheme which specifies all or most acceptable responses to each item.

communicative language competences
Competences which 'empower a person to act using specifically linguistic means' (Council of Europe, 2001, p. 9): linguistic, sociolinguistic and pragmatic competences.

comparative judgement
A way of breaking down judgement decisions to dichotomous ratings: a rater is presented with two performance samples and only has to decide which one is 'better.'

competence
The knowledge and ability to do something. In language testing, general competences and communicative language competences (linguistic, sociolinguistic, pragmatic) may play a role. See also: **skill**

component
Part of an examination, often competence-based (e.g. oral comprehension, reading comprehension). Components can also include integrated assessment (e.g. reading comprehension-into-written production). Also referred to as: section, part or paper.

Computer Adaptive Testing (CAT)
See: **adaptive testing**

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

constructed response
A form of written response to a test item that involves active production, rather than just choosing from a number of options (e.g. sentence-level answers, emails, essays).

construct-irrelevant variance
Factors unrelated to the tested skills which might affect the interpretation of scores (e.g. bias, poorly constructed items).

context expert
Someone who does not work in a particular domain, but has knowledge about it, especially related to Language Tests for Special Purposes (LSP). They can function as informant about specific language needs related to the domain, consultants about intercultural differences, item writers, etc. For example, LSP teachers for a language test of admission in the job.

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.

criterion-referenced test
A test in which the test taker's performance is interpreted in relation to predetermined criteria. Emphasis is on attainment of objectives rather than on test takers' scores as a reflection of their ranking within the group. Compare: **norm-referenced test**

cut-off score
The minimum score a test taker has to achieve in order to get a given grade in a test or an examination. In mastery testing, the score on a test which is considered to be the level required in order to be considered minimally competent or at 'mastery' level. Also known as *cut-off points, cut score*.

data integrity
Accuracy and completeness of digitally recorded test data. This needs to be checked

at the end of the test-taking process, making sure that the data is available and unaltered for further processing in the system (e.g. for marking or rating).

delivery
The stage in the assessment process where the test is logistically prepared and administered to test takers. Some tests have a fixed date of delivery (e.g. several times a year), others may be administered on demand (e.g. flexible dates), while others are permanently available.

delivery mode
Modality of administering the test, for example on paper or digital; in a testing centre or remote.

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.

design
Stage in the test development process in which the initial test specifications and sample test materials are produced and first decisions on marking and grading are taken.

Differential Item Functioning (DIF)
Property of an item to be less (or more) difficult for a subgroup of the test population without this being accounted for by a difference in ability.

digital skills
The test taker's ability to use digital technology effectively.

digital testing
Mode of administration of tests by use of devices like computers, laptops, tablets, etc. Traditionally called computer-based testing.
Compare: **paper-based testing**

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.

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/double rating
A method of assessing performance in which two markers or raters assess test taker performance on a test. The markers/raters can be human or automated systems.

effect size
A measure of the meaningfulness of a research outcome, in contrast to significance which just measures to which extent a result may be due to chance, and is dependent on sample size. If the sample used is large enough, even very small effects may be significant, although they have no practical impact.

embedded pretesting
When pretest items are disseminated within a test during its administration to get information from real test takers in a real test situation. Such items usually do not contribute to the test score. With this method, item data will not be distorted by inattention or excessive guessing, because test takers will solve them as they solve the rest of the items.

(standard) error of measurement
An estimate of the imprecision of a measurement. For example, if the error of measurement is 2, a test taker 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.

EU AI Act
Legislation governing the use of artificial intelligence within the European Union, including in testing contexts

examination
See: **test**

examine
Putting in practice the test construct and specifications of the test (delivery, marking and rating) by way of administering a concrete test version. The corresponding stages can be executed by humans with or without the support of technology and AI tools.

examiner
Anyone involved with putting in practice/executing the parts of the test which concern the interaction between the test and the test taker and the assessment part (test delivery of written and oral parts of the test; marking and rating). The term might be used in different testing organisations with

a different sense, sometimes covering only part of the tasks above.

fairness
A principle ensuring that test content and procedures do not disadvantage any group of test takers and that results are interpreted and used in an unbiased manner. It needs to be considered at the level of test design, administration and impact of results on individuals and groups. Compare: **justice**

fixed item selection
A test design where all test takers receive the same items.

focus group
A qualitative research method involving a small group of persons with similar characteristics (e.g. specialisation in a particular domain, working in the same environment, belonging to the same category of test takers, etc.). They engage in discussions about the chosen topic under the guidance of a moderator and offer in-depth information, opinions and perspectives on the topic in question.

formative assessment
Testing which takes place during, rather than at the end of, a course or programme of instruction. The results may enable the teacher to give remedial help at an early stage or change the emphasis of a course if required. Results may also help a student to identify and focus on areas of to be improved.

GDPR
See: **General Data Protection Regulation (GDPR)**

general competences
Competences which are not strictly related to language: knowledge of the world, sociocultural competence, intercultural competence, professional experience if any (Council of Europe, 2001, Section 5.1); savoir, savoir-faire, savoir-être, savoir apprendre (Council of Europe, 2020, Section 2.4).

General Data Protection Regulation (GDPR)
Regulation adopted at the level of the European Union concerning the protection of natural persons with regard to the processing of personal data and on the free movement of such data. Directly binding law for member states of the European Union.

generalise
Generalising from test takers' performances is 'drawing inferences about the test taker's ability beyond the immediate testing context' (McNamara, 2006).

generic scale
A scale that can be used independent of the task (as opposed to a task-specific scale, which refers to fulfilment of aspects of the concrete task, e.g. content points).

genuineness
The property of an input text or recording of being taken from real-life uses, as opposed to constructed to simulate them, for example using an excerpt from a news interview rather than writing a script for a simulated news interview and recording it, or re-recording the text of the original interview. 'Genuineness is a characteristic of the passage [the text] itself and is an absolute quality. Authenticity is a characteristic of the relationship between the passage and the reader and it has to do with appropriate response' (Widdowson, 1978, p. 80). Genuine texts can be produced by humans or generated by algorithms. A text can be genuine without being authentic in relation to a specific context and/or the interest or needs of the language user. At the same time, a text can be authentic without being genuine if it was edited for adaptation to the task. Compare: **authenticity**

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.

high-stakes language test
Language assessments with major consequences for the test taker's future (e.g. for obtaining citizenship, for university admission, for obtaining a job).

holistic scale
An instrument of assessment used for rating written and spoken productions consisting of bands to describe each level of performance, usually in general terms. The rater gives a single mark according to the general impression of the language produced, rather than by breaking it down into a number of marks for various aspects of language use. Compare: **analytic scale**

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

in-person test delivery
Test delivery organised in a location where the test takers go and sit the examination in the presence of proctors. Also known as *face-to-face test delivery*. Compare: **remote test delivery**

informant
Stakeholder who can help test developers understand test requirements in target language use (TLU) and the context in which the test will be administered. For example, for a test for nurses, these groups might be: nurses already working in the system; teaching nurses in LSP classes; doctors; health regulators and policy makers, etc.

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.

integrated task
A task assessing multiple language skills simultaneously, e.g. listening to writing: listening to a lecture and summarising it; reading to speaking: reading a (fragment of) a study and commenting on it as part of an oral presentation.

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

internal consistency
The degree to which a test or a subtest, or the ratings by different raters, measure the same latent variable. It is often assessed by Cronbach's alpha, a statistic which is based on the pairwise correlation of all items in the test.

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 or online. Also known as *proctor*.

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, employing either classical statistical indices such as facility and discrimination, or IRT indices.

item bank
A place to store test items and tasks, either on paper or digitally. The item bank may contain: item content; data on item use; data on item quality like difficulty and discrimination; and other metadata coding the items according to the test's requirements, e.g. topic, length of recording, number of speakers, etc.

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.

item review
A stage in the item and task production process, following piloting or pretesting. Evidence from this stage is used to retain, improve or reject items. Review can also be done in other stages of item and task production – for example, after they were written, a small team of experts might check them and provide feedback.

item type
Test items are referred to by names which tend to be descriptive of the form they take (e.g. multiple choice, sentence transformation, short answer, open cloze). Also known as *item format*.

item writer
A person preparing test content, such as input texts, items, tasks, with or without the contribution of AI.

item writing
The process of preparing test content, such as input texts, items, tasks, with or without the contribution of AI.

justice
The ethical and responsible use of language test results, especially in relation to the impact this might have on individuals and at the level of society. It concerns the values involved in the test construct and it involves monitoring the use of test results so as to promote equity, inclusion and social justice. Compare: **fairness**

Knowledge of Society (KoS) test
A test usually administered in the context of migration and integration, with the purpose of obtaining residency or citizenship in a country. The content may cover knowledge related to social, cultural, historical and legislative aspects in that country. The tests are normally provided in the majority language, which might introduce construct-irrelevant variance since the tests thereby become implicit language tests.

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

level
The degree of proficiency of a learner or test taker. According to the Common European Framework of Reference for Languages (CEFR) the levels are: A1 and A2 (basic user); B1 and B2 (independent user); C1 and C2 (proficient user). With the CEFR Companion volume (Council of Europe, 2020), Level pre-A1 was added. Other modalities of characterising the proficiency levels are: 'elementary', 'intermediate', 'advanced', etc.

linear-on-the-fly (LOFT) test construction
Dynamic test version construction during administration, by selection of pre-calibrated items from an item pool. It ensures that all test versions are different but comparable in content and difficulty. Sometimes also known as *on-the-fly test construction*.

linking of versions
A procedure which 'translates' results from one test or test version 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
A test at the time of its delivery with the purpose of measuring language skills and providing formal results, as opposed to pretesting or practice test (simulation).

logit
A logit is the unit of measurement used in IRT/Rasch analysis and Many-facet Rasch Measurement (MFRM).

low-stakes language test
A test with results which do not have a great impact on the test taker's future – for example, a classroom test to verify progress or a diagnostic test meant to identify a learner's strengths and weaknesses in a language.

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 or most 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. Also known as *answer key*.

marking
The assignment of a score to a test taker's response. May be used to refer specifically to cases in which the assignment is objective (for selected-response items or short-answer items with predefined correct responses), to distinguish this situation from rating.

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.

(item) metadata
Information on the item as it is included in the item bank (item type, difficulty level, topic, etc.). The information might vary according to type of test, mode of administration, etc. It is relevant for the construction of new test versions – appropriate items will be selected according to the metadata.

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.

monitoring
Observing, checking the process of test development and administration. For example, test performance can be monitored by observing and analysing the response of the test to the stakeholder needs over a medium and long term, through statistical data analysis, collecting and interpreting feedback, observing impact and score interpretation in relation to test takers' interests, etc.

needs analysis
A systematic gathering of specific information about the language needs of learners and the analysis of this information for purposes of language syllabus design. The procedures of needs analysis can be adapted to the activity of identifying [test] tasks and might involve the following steps: 1) identify the stakeholders who are familiar with relevant language use situations, who can help identify the relevant domain and tasks; 2) identify or develop procedures for gathering information about tasks; 3) gather information on the domain and tasks in collaboration with stakeholders; 4) analyse the tasks in terms of their task characteristics; and 5) make an initial grouping of tasks into categories of tasks with similar characteristics (Bachman & Palmer, 1996, p.102).

norm-referenced test
A test where scores are interpreted with reference to the performance of a given group, consisting of people comparable to the individuals taking the test. The term tends to be used of tests whose interpretation focuses on ranking individuals relative to the norm group or to each other. Compare: **criterion-referenced test**

objectively marked
An item which can be scored by applying a mark scheme, which contains all acceptable responses. They can be marked by human markers or by automated systems. Compare: **subjectively marked**

online platform
A digital system for administering and managing the test remotely.

online proctoring
Supervision of remote exams via tools like webcams and screen monitoring. Also known as *remote proctoring*.

open-ended question
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.

paper-based testing
Traditional mode of test delivery, by use of printed materials. Compare: **digital testing**

parallel versions
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. IRT can establish the parallelity of versions even without administering them to the same persons. Also known as *parallel or alternate forms*.

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 answer key.

partial qualifications
A test may include components for certain skills and not for others (e.g. only receptive skills or only oral skills). This structure is appropriate when 'only a more restricted knowledge of a language is required (e.g. for understanding rather than speaking)' (Council of Europe, 2001, 1.1).

piloting
Trying out test materials on a very small scale, perhaps by asking colleagues to respond to the items and comment in order to locate problems before launching a full-scale trial or pretest. Piloting is typically done on new types of items or tasks, or on constructed-response tasks where pretesting or trialling may not be possible.

planning
First phase in the process of specifications development in which information related to the characteristics of the test takers, the purpose of the test, the context, etc. is gathered from different informants and stakeholders for use in later stages.

point biserial correlation
A correlation coefficient that is used when one of the two variables to be compared is dichotomous.

pool
A group of items from which test versions may be selected; typically used for linear-on-the-fly tests (LOFTs) or adaptive tests.

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

practice test
A sample or training test for test taker familiarisation. Compare: **live test**

pre-entry test
A language test administered to migrants or refugees before they are allowed to enter the new country. The purpose of entry is usually family reunification or other processes related to immigration. The test is intended to facilitate integration in the new country.

pretesting
Administering the test to a number of test takers that will allow for statistical analysis, by way of checking its quality before it is put to use. Compare: **piloting** and **trialling**

proctor
Person who has responsibility to oversee the administration of an exam in the exam room or online. Also known as *invigilator*.

proctoring
Supervision during in-presence or remote test sessions to maintain order and detect irregularities. Also known as *invigilation*.

proficiency
Knowledge of a language and degree of skill in using it.

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

psychometric analysis
Statistical methods to evaluate test quality, reliability, and fairness.

qualitative analysis
The use of descriptive and interpretive methods (e.g. conversation analysis, verbal reports) to analyse data collected from individuals involved in test taking.

quality assurance
A system of processes and procedures designed to ensure that quality standards are met, and that assessment materials are effective and fit for the intended purpose.

quality control
Detailed procedures to detect and correct errors, for example an item review stage with specific checklists.

quantitative analysis
The use of statistical methods to analyse numerical data collected from language tests, like test scores, response times, frequency of linguistic features, etc.

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 (IRT).

rater
A person or an AI instance assigning marks to a test taker's performance by applying rating criteria. The terms 'marker', 'rater' and 'examiner' overlap and may be used in slightly differing ways by different test providers.

rater monitoring
The ongoing process of observing, evaluating, and providing feedback to raters to ensure that their scoring practices remain accurate, consistent, and in alignment with the rating instruments (scales, checklists, etc.). Monitoring also extends over rating performed by automated systems.

rating
Assigning a score to a test taker's written or oral productions using rating criteria such as task achievement, coherence, fluency, complexity, accuracy, etc. It can be done by human raters or by automated systems trained for the process.

rating criteria
The specific aspects of performance that are evaluated in written or oral productions, such as task achievement, coherence, fluency, complexity, accuracy, etc.

rating scale
A scale consisting of several ranked categories used for making subjective judgements by human assessors or trained judgements by automated systems. 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 scaling.

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

registration process
The method by which test takers enroll for a test and provide background data.

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.

remote proctoring
Supervision of remote exams via tools like webcams and screen monitoring. Also known as *online proctoring*.

remote test delivery
Online test taking from a non-central location (e.g. at home). Compare: **in-person test delivery**

reporting of results
Providing the test taker and any other stakeholders with the test results and any other information necessary to interpret them and to use them appropriately.

response
The test taker behaviour elicited by the input of a test. The written response can be selected (e.g. multiple choice, matching or ordering) or constructed (e.g. short responses or extended writing). If the test has a speaking component, the test taker's response will be in oral form. See also: **selected response** and **constructed response**

review
Evaluation of any element of a test, for example an item review or specifications review. The result of a review might or might not result in revision.

revision
Implementing changes in relation with different elements of the test (e.g. item revision, specification revision). Revision can come as a result of a review.

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

scaling
The process of applying statistical techniques to test taker responses to items to convert raw scores to a scale better suited for score reporting. Scaled scores 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. Scaling also ensures that scores from one test version can be directly compared to scores from another test version.

score
The total number of points someone achieves in a test, either before scaling (raw score) or after (scaled score).

score user
The institution, company, organisation which uses the test results for admission, promotion, qualification of the test taker.

scoring
The assignment of a numerical value to a response. It can happen as clerical marking of objective items or rating of productive tasks. It can be done by human assessors, by automated instruments or a combination of the two.

script
1. The test taker's responses to a test. The term is used particularly with open-ended task types.
2. A text that an actor reads (as in input texts for oral comprehension).

selected response
A form of response to a test item that involves choosing from a number of alternatives, rather than supplying an answer.

session
A scheduled occurrence of a test.

session history
A log of all test taker and system actions during a test session.

session recording
Capturing audio, video, and screen activity during (remote) tests.

skill
Ability to do something. Traditionally, in language testing the four skills of reading, writing, speaking and listening are often distinguished. See: **competence**

special needs
Visual, listening, speaking, learning, mobility difficulties of test takers which need to be addressed by the test taking organisation (testing centre). Accommodations must be arranged for the test takers with special needs so that they can have the proper conditions for fair and accurate assessment of their language abilities.

(test) specifications
A detailed set of documentation giving details of the test design, content, level, the types of task and item used, etc. They are written in accordance with the target population, target language use (TLU), intended use of the test, mode of administration, etc. The initial specifications are normally drawn up during the process of designing a new test or revising an existing one and are tested during the try-out phase. After try-out, the final specifications are written and the test is constructed based on them. The specifications may be part of test revision and be changed according to stakeholders' input and needs.

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

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
The term has different meanings depending on context. In statistics, the term typically has to do with values defined with reference to distance from an arithmetic mean ('standard deviation' and 'standard error of measurement'). Outside of statistics, the term typically means an expected level of performance, either on the part of test takers (as in 'standard setting') or for aspects of the testing programme (as in 'quality standards'). The term 'standardised' refers to tests or processes that adhere to procedural and/or quality standards.

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
Responses which must be scored using a rating scale or a similar instrument and which are not included as such in a mark scheme. They can be marked by human raters or by automated systems. Compare: **objectively marked**

system logging/tracking
Recording user actions and system events during a test session.

target language use (TLU) domain
A real-world context or area of life in which a language user needs to perform communicative tasks. It includes personal, public, occupational, and educational domains. In language testing, the TLU domain informs the design of tasks and content to ensure that test performance is relevant and generalisable to real-life use.

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.

task-specific scale
A rating scale which is tailored to the task which is rated by it, e.g. by enumerating content points that have to be covered in the answer.

test
Concept covering all manifestations of a tool for measuring proficiency represented by the test specifications: 'test' encompasses all possible test versions. A test will include different components (e.g. oral comprehension, reading comprehension, written production, written interaction). Also known as *examination*.

test administrator
A person responsible for setting up the test environment, managing logistics, and ensuring that testing procedures are correctly implemented.

test construction
The process of producing test versions, either in a fixed, repetitive structure or ad hoc, on-the-fly. Items and tasks for test construction can be selected from a bank or pool of materials. Human test producers usually select items and tasks for fixed structure tests, while digital systems select them for on-the-fly tests. Also known as *test assembling* or *test composition*.

test data analysis
Investigation and interpretation of qualitative and quantitative data related to a particular test. Different methods can be used and the purpose is the monitoring of: the way in which the test functions; if the results are used as intended by the test provider or the test sponsor; and what areas need to be improved.

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

test format
The structure of a test (e.g. test components, type and number of tasks and items), together with the way tasks and items are presented to the test taker (e.g. where and how answers are marked and how test takers navigate through the test).

test integrity
Fair and secure testing process, resulting in an uncompromised set of exam data (the response produced by the test taker) which can be further processed for issuing results. Measures need to be implemented and observed for securing test integrity.

test misuse
The use of a test beyond its original purpose and/or negative and harmful consequences of test (score) interpretation and use, irrespective of intentionality (Carlsen & Rocca, 2021).

test provider
The central organisation or entity responsible for the development, delivery, administration, scoring, and reporting of a language test.

test requirements
The test conditions which need to be met and features which need to be considered, as they result from the phase of planning in terms of level, degree of specificity, tested skills, target population, etc. They inform

the test design and need to be balanced with practical considerations before the test specifications are written.

test security
Measures to protect the confidentiality and integrity of test materials and results. It includes secure storage, invigilation/proctoring, detection techniques such as web crawling and statistical methods, and limiting access to test content.

test sponsor
The person or organisation who decides that a new test is necessary. They may commission the test to a testing organisation. Their role is to provide information about the target population and use of the test, to clarify the requirements of the test and to negotiate, together with the test developer, the balance between test requirements and practical considerations.

test taker
An individual who participates in the language assessment, taking a test. Also known as *candidate* or *examinee*.

test taker identification
Procedures to verify that the test taker is the correct registered person.

test template
A predefined exam structure that specifies the organisation, types, and sequence of tasks or questions to be delivered, serving as a reusable blueprint for generating individual test versions.

test usefulness
Initially defined by Bachman and Palmer, the concept requires that for any specific testing situation an appropriate balance must be achieved between certain qualities (e.g. validity, reliability, authenticity, interactivity, impact, practicality, fairness, quality of service) (Bachman & Palmer 1996; ALTE, 2020).

test version
The collection of test items and tasks administered to a test taker. It represents the implementation of the test specifications and differs from other test versions in the content of the items/tasks. It can be in a fixed form, on-the-fly or adaptive. Also known as *test form*.

Testing Cycle
The framework which includes all stages of test production, viewed as a circular process. This concept emphasises that the test production process is constantly monitored, reviewed and revised, with the results informing subsequent test administration sessions.

testing population
The group of individuals for whom a test is designed or to whom it is administered. The characteristics of the testing population such as age, linguistic background, proficiency level, cultural and educational context, and test-taking purpose are considered when designing a test.

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

traceability
The ability to review and reconstruct a test session which was administered in a digital format.

trialling
A stage in test development in which items or tasks (usually productive tasks such as essay writing) are tried out with a relatively small sample from the target population in order to determine whether the test functions as expected.

try-out
The phase of testing the draft specifications and making improvements based on practical experience and feedback from stakeholders, with the use of sample test items and tasks.

uneven profile
The (normal) case of a test taker (language user) of having higher levels of proficiency in some skills than in others (Carlsen, Rocca, & Machetti, 2023).

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. The terms 'validation' and 'validity' appear in multiple technical contexts throughout the manual. The comprehensive process of justifying test use is termed building a validity argument (Chapter 1). Specific procedures for confirming rating scales function appropriately are termed

scale validation (Chapter 2, Section 2.5). Statistical procedures for confirming measurement properties are psychometric validation (Appendix III). Data-based confirmation of theoretical predictions is termed empirical validation. While all contribute to the overall validity argument, they represent distinct methodological procedures with different requirements and outcomes.

validity argument
An exhaustive series of propositions and supporting evidence which seeks to substantiate the validity of interpretations recommended for test results.

verbal protocols (think-aloud protocols)
A qualitative analysis method based on data collected from test takers and/or examiners in which they talk about their thought processes while they take a test or assess a test performance (Banerjee, 2004).

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 next 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 1 of a test, Task 1 will account for a greater proportion of the total score than other tasks.

Bibliography and tools

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. AERA.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.

ALTE. (1994). *Code of practice*. Available online from the ALTE Resources – Free Guides and Reference Materials page (https://www.alte.org/Materials)

ALTE. (1998). *Multilingual glossary of language testing terms*. UCLES/Cambridge University Press. Available online from the ALTE Resources – Free Guides and Reference Materials page (https://www.alte.org/Materials)

ALTE. (2001a). *Development and descriptive checklist for tasks and examinations: General*. Available online from the ALTE Resources – Free Guides and Reference Materials page (https://www.alte.org/Materials)

ALTE. (2001b). *Individual component checklist: Reading*. Available online from the ALTE Resources – Free Guides and Reference Materials page (https://www.alte.org/Materials)

ALTE. (2001c). *Individual component checklist: Writing*. Available online from the ALTE Resources – Free Guides and Reference Materials page (https://www.alte.org/Materials)

ALTE. (2001d). *Individual component checklist: Listening*. Available online from the ALTE Resources – Free Guides and Reference Materials page (https://www.alte.org/Materials)

ALTE. (2001e). *Individual component checklist: Speaking*. Available online from the ALTE Resources – Free Guides and Reference Materials page (https://www.alte.org/Materials)

ALTE. (2001f). *Individual component checklist: Structural competence*. Available online from the ALTE Resources – Free Guides and Reference Materials page (https://www.alte.org/Materials)

ALTE. (2001g). *Individual component checklist – for use with one task: Reading*. Available online from the ALTE Resources – Free Guides and Reference Materials page (https://www.alte.org/Materials)

ALTE. (2001h). *Individual component checklist – for use with one task: Writing*. Available online from the ALTE Resources – Free Guides and Reference Materials page (https://www.alte.org/Materials)

ALTE. (2001i). *Individual component checklist – for use with one task: Listening*. Available online from the ALTE Resources – Free Guides and Reference Materials page (https://www.alte.org/Materials)

ALTE. (2001j). *Individual component checklist – for use with one task: Speaking*. Retrieved January 22, 2026, from the ALTE Resources – Free Guides and Reference Materials page (https://www.alte.org/Materials)

ALTE. (2001k). *Individual component checklist – for use with one task: Structural competence*. Available online from the ALTE Resources – Free Guides and Reference Materials page (https://www.alte.org/Materials)

ALTE. (2001l). *ALTE quality management and code of practice checklist 1: Test construction*. Available online from the ALTE Resources – Free Guides and Reference Materials page (https://www.alte.org/Materials)

ALTE. (2001m). *ALTE quality management and code of practice checklist 2: Administration and logistics*. Available online from the ALTE Resources – Free Guides and Reference Materials page (https://www.alte.org/Materials)

ALTE. (2001n). *ALTE quality management and code of practice checklist 3: Marking, grading, results*. Available online from the ALTE Resources – Free Guides and Reference Materials page (https://www.alte.org/Materials)

ALTE. (2001o). *ALTE quality management and code of practice checklist 4: Test analysis and post examination review*. Available online from the ALTE Resources – Free Guides and Reference Materials page (https://www.alte.org/Materials)

ALTE. (2002). *The ALTE Can Do Project*. Available online from the ALTE Resources – Free Guides and Reference Materials page (https://www.alte.org/Materials)

ALTE. (2005). *ALTE materials for the guidance of test item writers (1995, updated July 2005)*. Available online from the ALTE Resources – Free Guides and Reference Materials page (https://www.alte.org/Materials)

ALTE. (2014a). *The CEFR Grid for Speaking, developed by ALTE Members (input) v. 1.0*. Available online from the ALTE Resources – Free Guides and Reference Materials page (https://www.alte.org/Materials)

ALTE. (2014b). *The CEFR Grid for Writing Tasks v. 3.1 (analysis)*. Available online from the ALTE Resources – Free Guides and Reference Materials page (https://www.alte.org/Materials)

ALTE. (2014c). *The CEFR Grid for Writing Tasks v. 3.1 (presentation)*. Available online from the ALTE Resources – Free Guides and Reference Materials page (https://www.alte.org/Materials)

ALTE. (2016). *Language tests for access, integration, and citizenship: An outline for policy makers*. Association of Language Testers in Europe/Council of Europe. Available online from the ALTE Resources – Free Guides and Reference Materials page (https://www.alte.org/Materials)

ALTE. (2018). *Guidelines for the development of language for specific purposes tests: A supplement to the Manual for language test development and examining*. Available online from the ALTE Resources – Free Guides and Reference Materials page (https://www.alte.org/Materials)

ALTE. (2020). *Principles of Good Practice*. Association of Language Testers in Europe. Available online from the ALTE Resources – Free Guides and Reference Materials page (https://www.alte.org/Materials)

ALTE. (2023). *Uneven Language Profiles & Differentiated Language Requirements*. Available online from the ALTE Resources – Free Guides and Reference Materials page (https://www.alte.org/Materials)

ALTE. (2024a). *Association of Language Testers in Europe (ALTE) – Q-Mark*. Available online: https://www.alte.org/Setting-Standards

ALTE. (2024b). *Minimum standards for establishing quality profiles in ALTE examinations*. Available online from the ALTE Q-Mark page (https://www.alte.org/Setting-Standards)

Andrich, D., & Marais, I. (2019). *A Course in Rasch Measurement Theory: Measuring in the Educational, Social and Health Sciences*. Springer Link.

Assessment Systems. (2009). *Iteman 4* [Software]. Assessment Systems. Available online: https://assess.com/iteman/

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L. F. (2004). *Statistical analysis for language assessment*. 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., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

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

Bachman, L. F., Black, P., Frederiksen, J., Gelman, A., Glas, C. A. W., Hunt, E., … Wagner, R. K. (2003). Commentaries: Constructing an assessment use argument and supporting claims about test takers. *Measurement: Interdisciplinary Research & Perspective, 1*(1), 63–91.

Banerjee, J. (2004). *Reference supplement to the preliminary pilot version of the manual for relating language examinations to the Common European Framework of Reference for Languages: learning teaching and assessment. Section D: Qualitative analysis methods*. Available online: https://rm.coe.int/090000168092ac9c

Beacco, J.-C., & Porquier, R. (2007). *Niveau A1 pour le français. Un référentiel*. Éditions Didier.

Beacco, J.-C., & Porquier, R. (2008). *Niveau A2 pour le français. Un référentiel*. Éditions Didier.

Beacco, J.-C., Bouquet, S., & Porquier, R. (2004). *Niveau B2 pour le français. Un référentiel*. Éditions Didier.

Béland, S., & Falk, C. F. (2022). A comparison of modern and popular approaches to calculating reliability for dichotomously scored items. *Applied Psychological Measurement*, *46*(4), 321–337. https://doi.org/10.1177/01466216221084210

Béland, S., Cousineau, D. & Loye, N. (2017). Utiliser le coefficient omega de McDonald à la place de l'alpha de Cronbach. *McGill Journal of Education / Revue des sciences de l'éducation de McGill*, *52*(3), 791–804. https://doi.org/10.7202/1050915ar

Bloom, B. S. (1956). *Taxonomy of Educational Objectives: The Classification of Educational Goals: Handbook I Cognitive Domain*. Longmans.

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

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Erlbaum.

Brennan, R. L. (1992). *Generalizability theory*. Available online: https://ncme.org/wp-content/uploads/2025/10/Module-14-Generalizability-Theory-Brennan-Winter-1.pdf

Briggs, D. C. (2004). Comment: Making an argument for design validity before interpretive validity. *Measurement: Interdisciplinary Research & Perspective, 2*(3), 171–191.

British Council, ALTE, EALTA, & UKALTA. (2022). *Aligning Language Education with the CEFR: A handbook*. Available online: https://ealta.eu/documents/resources/CEFR%20alignment%20handbook.pdf

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Sage.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1–47.

Canale, M., & Swain, M. (1981). A theoretical framework for communicative competence. In A. S. Palmer, P. J. Groot, & S. A. Trosper (Eds.), *The construct validation of tests of communicative competence* (pp. 31–26). TESOL.

Carr, N. T. (2008). Using Microsoft Excel® to calculate descriptive statistics and create graphs. *Language Assessment Quarterly, 5*(1), 43–62. https://doi.org/10.1080/15434300701776336

CEFTrain. (2005). *CEFTrain*. Available online: https://blogs.helsinki.fi/ceftrain/

Chapelle, C.A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, *19*, 254–272. doi:10.1017/S0267190599190135

Chapelle, C. A., & Voss, E. (2021). *Validity argument in language testing: Case studies of validation research*. Cambridge University Press.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2007). *Building a validity argument for the Test of English as a Foreign Language*. Routledge.

CIEP. (2009). *Productions orales illustrant les 6 niveaux du Cadre européen commun de référence pour les langues*. Available online: https://www.france-education-international.fr/productions-orales-illustrant-les-6-niveaux-du-cecrl

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

Cizek, G. J. (1996). Standard-setting guidelines. *Educational Measurement: Issues and Practices*, *15*(1), 13–21.

Cizek, G. J., Ed. (2001). *Setting performance standards: Concepts, methods, and perspectives*. Erlbaum.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Sage.

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). *Setting performance standards: Contemporary methods*. Available online: http://www.edmeasurement.net/8225/Cizek-Bunch-Koons-2004-ITEMS.pdf

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, *17*(1), 31–44.

Cook, L., & Eignor, D. R. (1991). IRT equating methods. *Educational Measurement: Issues and Practice*, *10*(3), 37–45.

Corrigan, M. (2005). *Seminar to calibrate examples of spoken performance. Università per Stranieri di Perugia, CVCL (Centro per la Valutazione e la Certificazione Linguistica), 17–18 December 2005*. Available online: https://rm.coe.int/168045a0c9

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. Available online: http://www.coe.int/t/dg4/linguistic/Source/SourceForum07/D-Coste_Contextualise_EN.doc

Coulacoglou, C., & Saklofske, D. H. (2017). *Psychometrics and Psychological Assessment: Principles and Applications*. Academic Press.

Council of Europe. (1996). *Users' guide for examiners*. Language Policy Division.

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

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

Council of Europe. (2004a). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Available online: http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf

Council of Europe. (2004b). *Reference Supplement to the Preliminary Pilot version of the Manual for Relating Language examinations to the Common European Framework of Reference for Languages: learning, teaching assessment. Section D: Qualitative Analysis Methods*. Language Policy Division.

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]. Council of Europe.

Council of Europe. (2006a). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment. Writing samples*. Available online: https://www.eaquals.org/wp-content/uploads/Written-samples-complete_all-languages.pdf

Council of Europe. (2006b). *TestDaF sample test tasks*. Available online: https://rm.coe.int/168045a0d0

Council of Europe. (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR) – A manual*. Available online: http://www.coe.int/t/dg4/linguistic/Source/Manual%20Revision%20-%20proofread%20-%20FINAL.pdf

Council of Europe, & CIEP. (2009). *Productions orales illustrant les 6 niveaux du Cadre européen commun de référence pour les langues* [DVD]. Council of Europe/CIEP.

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*. Council of Europe Publishing.

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

Davies, A. (1997). (Guest Ed.). Ethics in language testing. *Language Testing, 14*(3). Available online: https://journals.sagepub.com/toc/ltja/14/3

Davies, A. (2004). (Guest Ed.). *Language Assessment Quarterly, 2 & 3*. Available online: https://www.tandfonline.com/journals/hlaq20

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

Davies, A., et al. (1999). *Dictionary of language testing*. UCLES/Cambridge University Press.

Dionne, E. & Béland, S. (2023). *Appliquer le modèle de Rasch: Défis et pistes de solution*. Presses de l'Université du Québec.

Dirkx, K. J. H., Kester, L., & Kirschner, P. A. (2014). The testing effect for learning principles and procedures from texts. *Journal of Educational Research, 107*(5), 357–364.

Downing, S. M., & Haladyna, T. M. (2006). *Handbook of test development*. Lawrence Erlbaum.

EALTA. (2006). *EALTA guidelines for good practice in language testing and assessment*. Available online: http://www.ealta.eu.org/documents/archive/guidelines/English.pdf

Eckes, T. (2009). *Reference supplement to the Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: learning teaching and assessment. Section H: Many-Facet Rasch Measurement*. Available online: https://rm.coe.int/1680459faa

Educational Testing Service. (2002). *ETS standards for quality and fairness*. ETS.

Embretson, S. E. (2007). Construct validity: A universal validity system or just another test evaluation procedure? *Educational Researcher, 36*(8). https://doi.org/10.3102/0013189X07311600

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Erlbaum.

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

Europarat., Council for Cultural Co-operation, Education Committee, Modern Languages Division., Goethe-Institut Inter Nationes., et al. (2001). *Gemeinsamer europäischer Referenzrahmen für Sprachen: Lernen, lehren, beurteilen*. Langenscheidt.

Ferne, T., & Rupp, A. A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly*, *4*(2), 113–148. doi: 10.1080/15434300701375923

Figueras, N., Kuijper, H., Tardieu, C., Nold, G., & Takala, S. (2005). *The Dutch Grid Reading/Listening*.

Figueras, N., & Noijons, J. (Eds.). (2009). *Linking to the CEFR levels: Research perspectives*. CITO/ Council of Europe. Available online: https://ealta.eu/documents/resources/Research_Colloquium_report.pdf

Frisbie, D. A. (1988). Reliability of scores from teacher-made tests. *Educational Measurement: Issues and Practice*, 7(1), 25–35. https://doi.org/10.1111/j.1745-3992.1988.tb00422.x

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment – An advanced resource book*. Routledge.

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

Glaboniat, M., Müller, M., Rusch, P., Schmitz, H., & Wertenschlag, L. (2005). *Profile Deutsch – Gemeinsamer europäischer Referenzrahmen. Lernzielbestimmungen, Kannbeschreibungen, Kommunikative Mittel, Niveau A1–C2*. Langenscheidt.

Goethe-Institut Inter Natione., KMK., EDK., & BMBWK. (eds.). (2001). *Gemeinsamer Europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen*. Langenscheidt.

Gorin, J. S. (2007). Reconsidering issues in validity theory. *Educational Researcher, 36*(8). Available online: 456–462. https://doi.org/10.3102/0013189X07311607

Green, R. (2013). *Statistical analyses for language testers*. Palgrave Macmillan. https://link.springer.com/book/10.1057/9781137018298

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]. Guerra Edizioni.

Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice, 18*(4), 5–9.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38–47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Harvill, L. M. (1991). Standard error of measurement. *Educational Measurement: Issues and Practice*, *10*(2), 33–41.
https://doi.org/10.1111/j.1745-3992.1991.tb00195.x

Heaton, J. B. (1990). *Classroom testing*. Longman.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). American Council on Education/Praeger.

Huang, L. (2021). Introducing integrated language skills assessment at the language department of a Czech university. *Language Learning in Higher Education*, *11*(1), 253–262.

Hughes, A. (1989). *Testing for language teachers*. Cambridge University Press.

ILTA. (2000). *ILTA code of ethics*. Available online: https://cdn.ymaws.com/www.iltaonline.com/resource/resmgr/docs/ilta_code_english.pdf

ILTA. (2007). *ILTA guidelines for practice*. Available online: https://cdn.ymaws.com/www.iltaonline.com/resource/resmgr/docs/ilta_guidelines.pdf

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

JCTP. (2004). *Code of fair testing practices in education*. Available online: https://www.apa.org/science/programs/testing/fair-testing.pdf

JLTA. (n.d.). *Code of good testing practice*. Available online: http://www.avis.ne.jp/~youichi/COP.html

Jones, N., & Saville, N. (2009). European language policy: Assessment, learning and the CEFR. *Annual Review of Applied Linguistics, 29*, 51–63.

Jones, P., Smith, R. W., & Talley, D. (2006). Developing test forms for small-scale achievement testing systems. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 487–525). Erlbaum.

Jones, R. L., & Tschirner, E. (2006). *A frequency dictionary of German – Core vocabulary for learners*. Routledge.

Kaftandjieva, F. (2004). *Reference supplement to the preliminary pilot version of the manual for relating language examinations to the Common European Framework of Reference for Languages: learning teaching and assessment. Section B: Standard setting*. Available online: https://rm.coe.int/1680459faa

Kane, M. (2002). Validating high stakes testing programs. *Educational Measurement: Issues and Practice, 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 R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Washington, DC: American Council on Education/Praeger.

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

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

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

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where the criteria should come from? *Assessing Writing, 16*(2), 81–96.

Kolen, M. J. (1988). Traditional equating methodology. *Educational Measurement: Issues and Practice, 7*(4), 29–37.

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 155–186). American Council on Education/Praeger.

Kuijper, H. (2003). *QMS as a continuous process of self-evaluation and quality improvement for testing bodies*.

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

Kunnan, A. J. (ed.). (2000b). *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida*. UCLES/Cambridge University Press.

Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. Weir (Eds.), *European language testing in a global context – Proceedings of the ALTE Barcelona Conference, July 2001*. UCLES/Cambridge University Press.

Lewkowicz, J. A. (2000). Authenticity in language testing: some unresolved issues. *Language Testing, 17*(1), 43–64.

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement, 3*, 103–122.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*(1), 85–106.

Linacre, J. M. (2009). *Facets 3.64.0* [Software]. https://www.winsteps.com/index.htm

Linacre, J. M. & Wright, B. D. (1987). *Item bias: Mantel Haenszel and the Rasch Model*. https://www.rasch.org/memo39.pdf

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, *36*(8), 437–448. https://doi.org/10.3102/0013189X07311286

Lissitz, R. W., & Samuelsen, K. (2007). Further clarification regarding validity and education. *Educational Researcher*, *36*(8), 482–484. https://doi.org/10.3102/0013189X07311612

Livingston, S. A. (2004). *Equating test scores (without IRT)*. Available online: http://www.ets.org/Media/Research/pdf/LIVINGSTON.pdf

Luoma, S. (2003). *Assessing speaking*. Cambridge University Press.

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods 42*, 847–862. doi: 10.3758/BRM.42.3.847

McCowan, R. J., & McCowan, S. C. (1999). *Item analysis for criterion-referenced tests*. Available online: https://files.eric.ed.gov/fulltext/ED501716.pdf

McNamara, T. (1996). *Measuring second language performance*. Addison Wesley Longman Ltd.

McNamara, T. (2006). Validity and values: Inferences and generalizability in language testing. In M. Chahloub-Deville, C. A. Chapelle, & P. Duff (Eds.), *Language learning & language teaching, Volume 12: Inference and generalizability in applied linguistics* (pp. 27–46). John Benjamins Publishing Co.

McNamara, T., & Roever, C. (2006a). Fairness reviews and codes of ethics. *Language Learning, 56*(S2), 129–148.

McNamara, T., & Roever, C. (2006b). *Language testing: The social dimension*. Blackwell.

McNamara, T., & Ryan, K. (2011). Fairness versus justice in language testing: The place of English literacy in the Australian citizenship test. *Language Assessment Quarterly, 8*(2), 161–178.

McNamara, T., Knoch, U. & Fan, J. (2019). *Justice and Language Assessment*. Oxford University Press.

Messick, S. (1989a). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.

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

Mislevy, R. J. (2007). Validity by design. *Educational Researcher*, *36*(8), 463–469. https://doi.org/10.3102/0013189X07311660

Mislevy, R. J., Steinberg, L. S., & 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), 470–476. https://doi.org/10.3102/0013189X07311608

North, B., & 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*. Available online: https://rm.coe.int/1680459fa0

O'Sullivan, B. (2020). *The Comprehensive Learning System*. British Council. Available online: https://www.britishcouncil.org/sites/default/files/cls_bcps1_bos_30-09-2020_final.pdf

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

Perlmann-Balme, M., & Kiefer, P. (2004). *Start Deutsch: Deutschprüfungen für Erwachsene. Prüfungsziele, Testbeschreibung*. Goethe-Institut/WBT.

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

Piccardo, E., Berchoud, M., Cignatta, T., Mentz, O., & Pamula, M. (2011). *Pathways through assessing, learning and teaching in the CEFR*. Available online: https://www.ecml.at/Portals/1/documents/ECML-resources/2011_08_29_ECEP_EN_web.pdf

Ross, S. J., & Okabe, J. (2009). The subjective and objective interface of bias detection on language tests. *International Journal of Testing, 6*(3), 229–253. https://doi.org/10.1207/s15327574ijt0603_2

Saville, N. (2005). Setting and monitoring professional standards: A QMS approach. *Research Notes, 22*, 2–5.

Schneider, G., & Lauber, M. (2020). *Statistics for linguists: A patient, slow-paced introduction to statistics and to the programming language R*. University of Zurich. https://dlf.uzh.ch/openbooks/statisticsforlinguists/

Shermis, M. D., & Wilson, J. (eds.). (2024). *The Routledge International Handbook of Automated Essay Evaluation*. Routledge.

Sireci, S. G. (2007). On Validity Theory and Test Validation. *Educational Researcher*, *36*(8), 477–481. https://doi.org/10.3102/0013189X07311609

Sickinger, R., Brunfaut, T. & Pill, J. (2025). Comparative judgement for evaluating young learners' EFL writing performances: Reliability and teacher perceptions of holistic and dimension-based judgements. *Language Testing, 42*(2), 137–166.

Sijtsma, K. (2008). On the use, the misuse, and the very limited usefulness of Cronbach's Alpha. *Psychometrika*, *74*(1), 107–120. doi: 10.1007/s11336-008-9101-0

Spinelli, B., & Parizzi, F. (2010). *Profilo della lingua italiana. Livelli di riferimento del QCER A1, A2, B1 e B2*. RCS Libri – Divisione Education.

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

Stiggins, R. J. (1987). *Educational Measurement: Issues and Practice*, *6*(3), 33–42. https://doi.org/10.1111/j.1745-3992.1987.tb00507.x

Taylor, L. (Ed.). (2011). *Examining speaking: Research and practice in assessing second language speaking*. UCLES/Cambridge University Press.

Taylor, L. & Banerjee, J. (2023). Accommodations in language testing and assessment: Safeguarding equity, access and inclusion. *Language Testing, 40*(4), 847–855. https://doi.org/10.1177/02655322231186221

Traub, R.E., & Rowley, G.L. (1991). Understanding reliability. *Educational Measurement: Issues and Practice*, *10*(1), 37–45.

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

Tsagari, D., & Spanoudis G. (eds.). (2013). *Assessing L2 Students with Learning and Other Disabilities*. Cambridge Scholars Publishing.

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]. 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]. University of Cambridge ESOL Examinations.

University of Cambridge ESOL Examinations & Council of Europe. (2009b). *Common European Framework of Reference for Languages: Examples of speaking performance at levels A2 to C2*. Available online: https://www.cambridgeenglish.org/Images/22649-rv-examples-of-speaking-performance.pdf

Van Avermaet, P. (2003). *QMS and the setting of minimum standards: Issues of contextualisation variation between the testing bodies*.

Van Avermaet, P., Kuijper, H., & 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., & Trim, J. (1990). *Waystage 1990*. Cambridge: Cambridge University Press.

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

Van Ek, J. A., & 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 Common European Framework of Reference for Languages: learning teaching and assessment. Section C: Classical Test Theory*. Available online: https://rm.coe.int/1680459faa

Verhelst, N. (2004b). *Reference supplement to the preliminary pilot version of the manual for relating language examinations to the Common European Framework of Reference for Languages: learning teaching and assessment. Section E: Generalizability Theory*. Available online: https://rm.coe.int/1680459faa

Verhelst, N. (2004c). *Reference supplement to the preliminary pilot version of the manual for relating language examinations to the Common European Framework of Reference for Languages: learning teaching and assessment. Section F: Factor Analysis*. Available online: https://rm.coe.int/1680459faa

Verhelst, N. (2004d). *Reference supplement to the preliminary pilot version of the manual for relating language examinations to the Common European Framework of Reference for Languages: learning teaching and assessment. Section G: Item Response Theory*. Available online: https://rm.coe.int/1680459faa

Ward, A. W., & Murray-Ward, M. (1994). *Guidelines for development of item banks (Instructional Topics in Educational Measurement Series 17)*. Available online: https://ncme.org/wp-content/uploads/2025/10/Module-17-Item-Bank-Development-Ward-Murray-War-1.pdf

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.

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

Weir, C. J., & Milanovic, M. (Eds.). (2003). *Continuity and innovation: Revising the Cambridge Proficiency in English Examination 1913–2002*. UCLES/Cambridge University Press.

Widdowson, H. G. (1978). *Language teaching as communication*. Oxford University Press.

Williamson, D.M., Xi, X., & Breyer, F.J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2–13.

Wright, B. D., & Stone, M. H. (1979). *Best Test Design: Rasch Measurement*. Mesa Press.

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

Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores*. Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zwick, R. (2012). *A Review of ETS Differential Item Functioning Assessment Procedures: Flagging Rules, Minimum Sample Size Requirements, and Criterion Refinement*. RR-12-08. ETS. Available online: https://files.eric.ed.gov/fulltext/EJ1109842.pdf

Acknowledgements

This Manual is a revised version of an earlier one, with the same title, produced by ALTE on behalf of the Language Policy Division, Council of Europe and published in 2011.

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 revision team would like to thank the ALTE Secretariat for their constant support in developing the project.

We would like to acknowledge the contributions of the special interest (SIG) group participants at the meetings in Bad Homburg, Istanbul, Cardiff and Pécs, who provided useful feedback and commentary.

The editing team for this revision:

Dina Vîlcu
Beate Zeidler
Mika Hoffman

Lead members of the revision team:

Mika Hoffman
Vincent Folny
Dina Vîlcu
Darren Perrett
Beate Zeidler
Dominique Casanova
Fabiana Pica
Marta Garcia
Bert Wylin

Members of the revision team:

Joaquín Cruz
Rita Marzoli
Seçil Alaca
Antonella Mastrogiovanni
Natasha Weeds
Marvin Möckel
Willemijn van den Berg
Tristan Logiewa
Marius Ullrich
Joaquín Cruz Trapero
Anthea Wilson
Marta Desimoni
Francesca la Russa

Council of Europe reviewers:

José Noijons
Gabor Szábo

ALTE reviewers:

Graham Seed
Waldemar Martyniuk
Letizia Cinganotto
Javier Fruns Gimenez
Lorenzo Rocca
Kateřina Vodičková
Stefanie Dengler
Sibylle Plassmann
Anna Mouti

Images 11, 12 and 13 in Chapter 5:

Thomas Sanders (Televic Education)

The publishing team:

John Savage

Cheshire Typesetting

Mariangela Marulli

Eleni Karagianni

The Council of Europe is the continent's leading human rights organisation. It comprises 46 member states, including all members of the European Union. All Council of Europe member states have signed up to the European Convention on Human Rights, a treaty designed to protect human rights, democracy and the rule of law. The European Court of Human Rights oversees the implementation of the Convention in the member states.

Produced by:
Association of Language Testers in Europe
Shaftesbury Road
Cambridge, CB2 8EA
United Kingdom
www.alte.org

On behalf of:
Council of Europe

ALTE Association of Language Testers in Europe

COUNCIL OF EUROPE
CONSEIL DE L'EUROPE