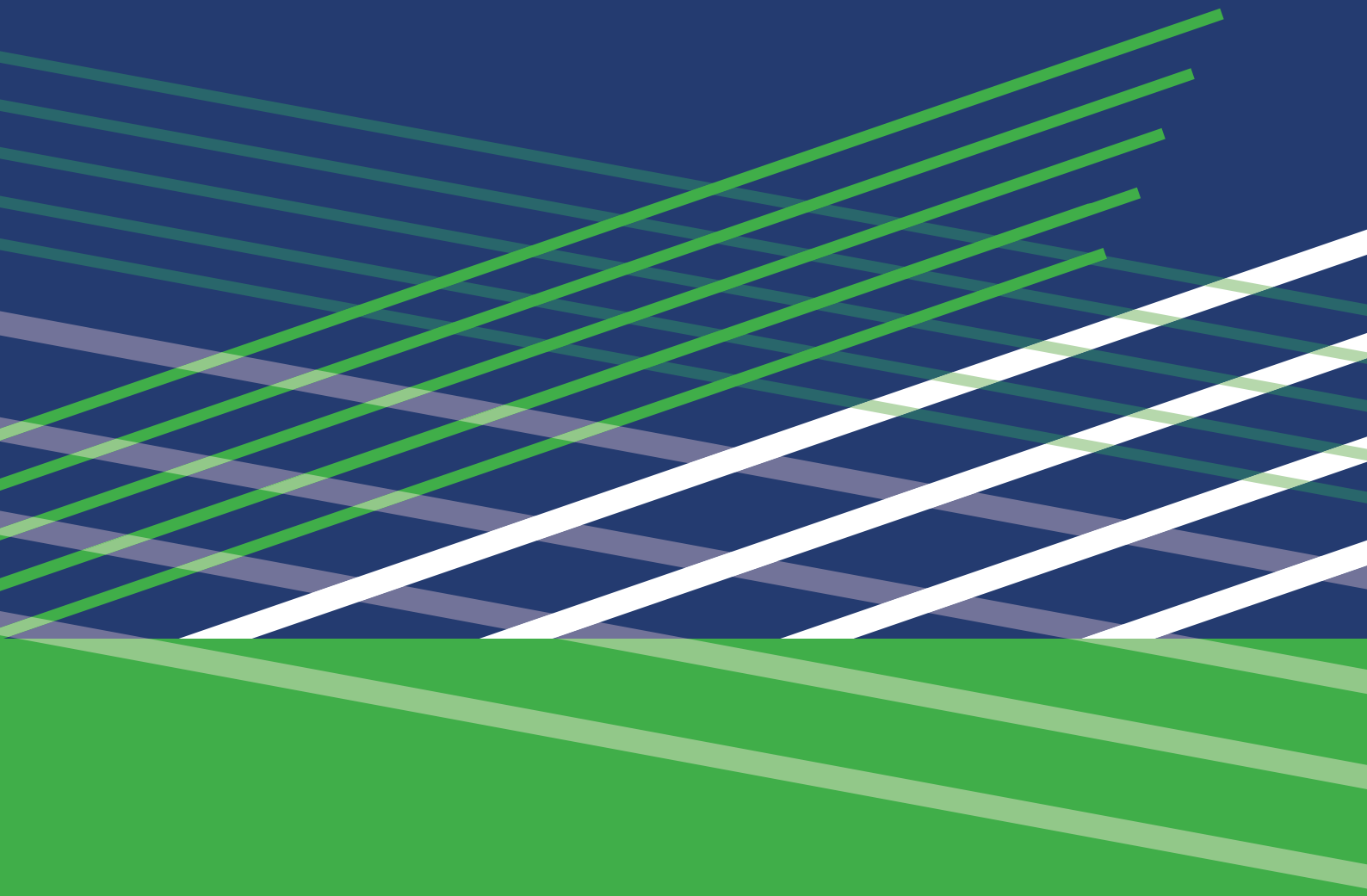
Association of Language Testers in Europe

ALTE

ALTE Principles of Good Practice



ALTE 2020

ALTE

Principles of Good Practice 2020

Acknowledgements

The Principles of Good Practice Working Group operated from 2015 to 2019 and consisted of the following individuals over the course of these years, who are acknowledged and thanked for their hard work:

Emyr Davies
Waldemar Martyniuk (chair)
Siuan Ni Mhaonaigh
Jose Pascoal
Michaela Perlmann-Balme
Nick Saville
Graham Seed
Cathy Taylor
Henna Tossavainen
Koen Van Gorp

Many other ALTE Members and Affiliates have reviewed and commented on the text, most notably during the sessions at the Salamanca meeting of November 2018. They are too many to name in full, but they are thanked for their input. In addition, there are numerous individuals who worked on previous versions of ALTE's Principles of Good Practice and Code of Practice documents over the last 25 years, whose efforts have built a foundation for the current document and they are likewise thanked.

Production Team:
George Hammond Design
Eleni Karagianni
Mariangela Marulli
Jane Rust
John Savage
Graham Seed

Contents

Foreword

2020 is an important milestone in the history of ALTE as it marks 30 years of collaboration between the Members. In 1990, the first exploratory meeting took place in Barcelona and was attended by representatives of assessment providers for eight different languages. In light of this successful event, the Association was officially formed a year later and collaboration has continued since then. As of January 2020, there are now 33 Full Members representing 25 languages and with many hundreds of Institutional and Individual Affiliates.

From the start, ALTE's ambitions were clearly stated: to bring together organisations of different types from across Europe to develop and set standards for language assessment, recognising the diversity of language learning and assessment practices and the need to exchange know-how and good practice to achieve the shared goals. This history of this multilingual enterprise has now been told in *The History of ALTE – The Association of Language Testers in Europe: The first 30 years*. A central part of this history was the development of ALTE's Principles of Good Practice. The first edition of the document was the result of an ALTE project during the early days of the Association that also produced ALTE's Code of Practice (1993/4).

This current document is the third and latest edition of the Principles. Minor revisions were made to the original during the 1990s and a second edition was published in 2001. Since then much has changed in the world, and so I am very pleased to introduce this new version. The Principles themselves have been thoroughly reviewed, updated and extended as a result of a lengthy consultation and editing process with Members and other stakeholders across ALTE networks (2015–19). Over the three decades of working together, the Membership of ALTE has grown and consequently a much wider community of practice was able to take part in the consultation process that fed into the revision of these latest Principles.

There are two distinctive features of the ALTE Principles of Good Practice (PoGP) that can be highlighted.

First, while drawing extensively from the literature in the wider field of assessment, the Principles have been designed specifically to serve the needs of the ALTE Members in developing language assessments for their own stakeholders. This means they have been written to be accessible to a wide audience with varying degrees of professional expertise and resources at their disposal. In this respect, the approach taken contributes to 'language assessment literacy' and helps ALTE Members and other participants to understand assessment better in their own contexts.

Secondly, the Principles are intended to fit into the ALTE Quality Management System (QMS) as an integral reference document to be used in setting and monitoring the ALTE Standards. It is used alongside the Quality Checklists and the other materials that form part of the Procedures for Auditing that lead to the awarding of the ALTE Q-mark.

The revision project itself was coordinated by a small working group that planned the collaborative activities and enabled the consultative activities to take place, both online and via workshops at meetings. Members of the working group drafted the revised texts and additional sections but the final document is the result of an extensive moderation process. It is thus endorsed by the membership as a whole.

Inevitably there will be improvements that can be made and feedback from users of the document is welcomed. In the spirit of continuous improvement, this feedback will be collected and used to inform further revisions – and so I look forward to the fourth edition sometime in the future.

Nick Saville
ALTE Secretary-General

About ALTE

The Association of Language Testers in Europe (ALTE) was set up as a result of a meeting in 1989 by the Universities of Cambridge and Salamanca. The initial aim was to establish common standards for language testing across Europe, thereby supporting multilingualism and helping preserve the rich linguistic heritage of Europe. It was also vital that individual test takers gained a language qualification that was a fair and accurate assessment of their linguistic ability, one which was recognised around the world, and which could be accurately compared to qualifications in other languages.

By 2020, ALTE expanded its membership to include 33 Full Members representing 25 European languages, as well as around 60 Institutional Affiliates and over 500 Individual Affiliates from all around the world (see here for the current numbers).

Our shared primary aims are to:

- establish quality standards and principles of good practice for all stages of the language testing process;
- promote transnational recognition of language certification;
- improve the quality of language assessment through joint projects, sharing best practice, and the work of special interest groups;
- provide training and enhance assessment literacy among language professionals and the wider community;
- raise awareness of issues relating to language testing through regular meetings;
- promote the benefits of plurilingualism/multilingualism and language learning;
- provide thought leadership through international conferences;
- achieve positive impact on educational processes and on society in general.

As international mobility increases, there is a growing need for transferable, comparable language qualifications which are meaningful to stakeholders/test users (including candidates, employers, policy makers and others), and which offer real career and social advantages to the individual. To meet this need, ALTE has established a set of 17 common quality standards for its Members' exams, which cover all stages of the language testing process: test development, task and item writing, test administration marking and grading, reporting of test results, test analysis and reporting of findings.

As a result, users of the exams – whether individuals, employers, educational institutions or government bodies – can be confident that the language assessments devised and delivered by ALTE Members meet specified professional standards.

The ALTE Q-mark is one such language testing quality indicator which Member organisations can use to show that their exams have passed a rigorous audit and meet all 17 of ALTE's quality standards. The Q-mark allows test users to be confident that an exam is backed by appropriate processes, criteria and standards.

Since 2018, ALTE Members have been engaged in a strategic review of the association and its future. This process resulted in a decision to a more flexible and inclusive approach for all by changing the formal status of the association from a European Economic Interest Grouping (EEIG), which restricts membership to EU/EEA, to a Charitable Incorporated Organisation (CIO), registered in England. This change, completed in 2019, will increase the geographical scope of ALTE activities and membership, and allow the association to be open to new opportunities.

The ALTE Principles of Good Practice

The initial document entitled *Principles of Good Practice for ALTE Examinations* was drafted by Nick Saville and Mike Milanovic between 1991 and 1993 and discussed at several ALTE meetings (Alcalà de Henares, 1992, Paris and Munich, 1993). The document was intended to set out in more detail the principles which ALTE Members should adopt in order to achieve their goals of high professional standards. The approach to achieving good practice was influenced by a number of sources from within the ALTE membership and from the field of assessment at large (e.g. the work of Lyle Bachman, Samuel Messick (1980) and the American Educational Research Association/American Psychological Association/National Council on Measurement in Education (AERA/APA/NCME) *Standards for educational and psychological testing*, 1985). ALTE Members sought feedback on the document from eminent external experts in the field (Bernard Spolsky, Lyle Bachman). While it was not published in its entirety, parts of the document were later incorporated into the Users' Guide for Examiners produced by ALTE on behalf of the Council of Europe (published 2011, in a revised version, as the *Manual for Language Test Development and Examining*). In 1994 the ALTE Code of Practice was adopted, intended to be a general statement of what the users of the examinations should expect and of the roles and responsibilities of stakeholders in striving for fairness. The ALTE Code of Practice Working Group was set up in 2000 and used the Code as a basis for the specification of a set of 17 minimum professional standards (Minimum Standards, MS) that became the central common reference for the ALTE Quality Management System (QMS) (see Chapter 3 for details).

In 2001, a revised version of *Principles of Good Practice for ALTE Examinations* was published on the ALTE website as a 'Revised Draft – 2001' in English and in Galician, and included in the set of ALTE resource documents. It addressed more specifically the central issues of validity and reliability and looked at the related issues surrounding the impact of examinations on individuals and on society. The 2001 version, like the earlier drafts, drew heavily on the *Standards* for educational and psychological testing (AERA/APA/NCME 1999) – especially in the sections on validity and reliability – as well as the work of Bachman (1991) and Bachman and Palmer (1996). In the fields of psychological and educational assessment, the USA has a long tradition of setting standards. In the USA since the early 1980s, Educational Testing Service (ETS) have produced their own Standards for Quality and Fairness (1981, 1987, 2000) drawing heavily on the AERA/APA/NCME Standards. The International Language Testing Association (ILTA) conducted a review of international testing standards in 1995 and in 2000 published its Code of Ethics (see Chapter 1).

In 2015, ALTE Members decided to set up a Working Group to review the PoGP. It was felt that although the principles remain as valid as they were when the document was first published, their presentation may need to be expanded by taking into consideration the latest developments in the field of language testing and assessment, such as the socio-cognitive approach, the ethical shift, and the growing Common European Framework of Reference for Languages (CEFR, first published in 2001) toolkit offered by the Council of Europe in close cooperation with ALTE. A need was also voiced to take a closer look at all the strategic documents adopted by ALTE in the last decades in order to clarify their status, function, and hierarchy, and revise if necessary – in other words to come up with a coherent set of documents that constitute the Association. The first meeting of the Working Group took place in November 2015 in Perugia. The review of the PoGP document was conducted in open consultation with all ALTE Members and offered an opportunity for a more strategic discussion on the mission and the function of the Association. The Working Group prepared a first draft of the new version of the ALTE PoGP – conceptualised as a coherent set of guidelines that all ALTE Members/Institutional Affiliates share and subscribe to – and invited the whole membership to contribute to the preparation of the final version to be published on the ALTE website. The aim of a plenary and the following workshops during the ALTE meeting in Salamanca in November 2018 was to present the draft of the document and invite feedback from the membership. Members/Institutional Affiliates were invited to register for one of the following four thematic areas corresponding to the four main sections of the new PoGP document:

- **Workshop 1:** Ethical Considerations and the ALTE Code of Practice
- **Workshop 2:** The Concept of Usefulness in Examinations and Examination Qualities
- **Workshop 3:** ALTE Quality Management System
- **Workshop 4:** ALTE Support and Resources

Each workshop was chaired by members of the PoGP Working Group who introduced the relevant sections and invited feedback. A Rapporteur was assigned in each of the four groups to take notes and presented a summary of the feedback in a plenary session following the workshops. The feedback of the ALTE Members and Institutional Affiliates was highly appreciated and taken into consideration in preparation of the final version of the PoGP document.

Introduction

All assessment has to take into consideration the language learner/user and the situations that he or she finds himself or herself in. In today's Europe, any language learner will find that the way in which they learn and use languages will be influenced by a number of various educational, social, affective and other factors which will impact the learner's cognition and language knowledge.

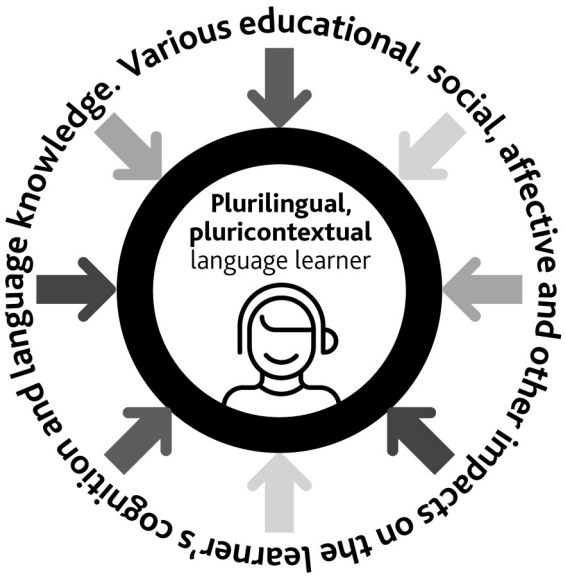


Figure 1: The plurilingual, pluricontextual language learner/user

Language learners/users will naturally have one, if not more, languages (which may include 'dialects') that they may call 'home languages', because they used them during their formative years. Throughout their life they will find themselves using language, in the broadest sense, as well as different languages, in varied contexts with many other users at different levels of competence, and they will acquire new language to bridge communicative gaps that exist between themselves and others. These interactions mean the language user becomes a plurilingual, pluricontextual language learner, and sites the acquisition of language very much within a socio-cognitive understanding.

As the learner develops language knowledge and communicative skills, there will be situations when these need to be assessed, perhaps to demonstrate to others in society that the relevant language has been acquired to be able to adequately take part in a particular context. In the context of 'learning-oriented assessment' (Jones and Saville 2016), assessments will be made during the process of learning, some of which may be performed in the learning environment and others as large-scale summative assessment.

In each assessment setting, the learner interacts with an assessment task, or tasks, and this interaction produces a performance. The observed performance on an assessment task produces feedback, which may be a score on a scale, or it may be some sort of more detailed feedback. Whatever the feedback, provided it is *meaningful*, it will impact the learner to understand his or her communicative competences to date, and may also guide the learner in terms of future ambitions of communicative competences. Furthermore, the feedback will also affect other test users and stakeholders, as well as society as a whole. This will also undoubtedly affect the learner again.

Assessment therefore has to be of sufficient quality to ensure that when a language learner encounters a test task, the observed performance results in feedback which is accurate and meaningful. Fairness is therefore an overriding concern in all aspects of assessment and provides a context for the principles of good practice. Whenever examinations are widely

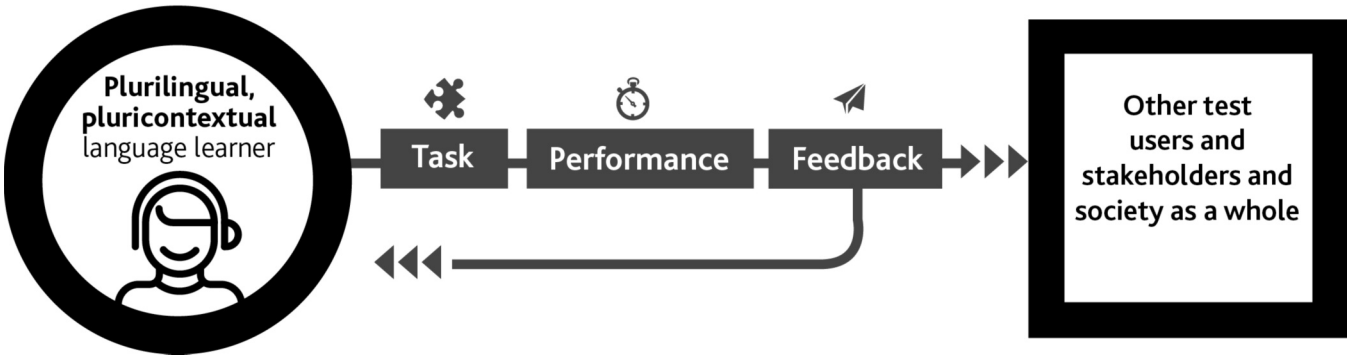


Figure 2: The role of language assessment

used within educational contexts they affect not only individuals but also institutions and society as a whole, as is demonstrated above. Given the potentially wide-ranging impact of examinations, it is important for both examination developers and users to implement standards which will ensure that the assessment procedures are of high quality and that all stakeholders are treated fairly. A code of practice of this kind must be based on sound principles of good practice in assessment which allow high standards of quality and fairness to be achieved.

The discussion of what constitutes good practice presented in this document is an attempt to reflect a concern for accountability in all areas of assessment which are undertaken by ALTE Members. It recognises the importance of validation through the collection of data and the role of research and development in examination processes. In this respect, it is likely that the principles which are outlined below will continue to evolve over time as research and development programmes expand.

1. Ethical considerations

In questions regarding ethics, ALTE expects its Members to apply the ILTA Code of Ethics. In addition, ALTE would like to highlight the importance of social ethics: social responsibility and justice in all assessment and in particular in using language tests for questions of migration and by political decision-makers. Social ethics would also entail interaction and cooperation between stakeholders. ALTE would further like to highlight the importance of making such ethical choices in using digital technology that would support learning.

1.1 The ILTA Code of Ethics

The ILTA Code of Ethics identifies nine fundamental principles which 'draw upon moral philosophy and serve to guide good professional conduct', each elaborated on by a number of annotations which clarify the nature of the principles[1].
As stated in the introductory notes, this Code of Ethics 'does not release language testers from the obligations and responsibilities laid on them by other Codes to which they have subscribed or from their duties under the legal codes, both national and international, to which they may be subject'.
Another general remark included in the introduction highlights the responsibility of the individuals involved in language testing to apply independent judgement in their actions:

> *Language testers are independent moral agents and sometimes they may have a personal moral stance which conflicts with participation in certain procedures. They are morally entitled to refuse to participate in procedures which would violate personal moral belief. Language testers accepting employment positions where they foresee they may be called on to be involved in situations at variance with their beliefs have a responsibility to acquaint their employer or prospective employer with this fact. Employers and colleagues have a responsibility to ensure that such language testers are not discriminated against in their workplace.*

1.2 Language Assessment for Migration and Integration

ALTE, as a long-standing partner for the Council of Europe, increased the collaboration with the Education Policy Division over the last decade, with particular regard to projects and research related to the context of migration. During the past

[1] For annotations to each of the principles see the original version of the document, available at: www.iltaonline.com

THE ILTA PRINCIPLES

Principle 1 Language testers shall have respect for the humanity and dignity of each of their test takers. They shall provide them with the best possible professional consideration and shall respect all persons' needs, values and cultures in the provision of their language testing service.

Principle 2 Language testers shall hold all information obtained in their professional capacity about their test takers in confidence and they shall use professional judgement in sharing such information.

Principle 3 Language testers should adhere to all relevant ethical principles embodied in national and international guidelines when undertaking any trial, experiment, treatment or other research activity.

Principle 4 Language testers shall not allow the misuse of their professional knowledge or skills, in so far as they are able.

Principle 5 Language testers shall continue to develop their professional knowledge, sharing this knowledge with colleagues and other language professionals.

Principle 6 Language testers shall share the responsibility of upholding the integrity of the language testing profession.

Principle 7 Language testers in their societal roles shall strive to improve the quality of language testing, assessment and teaching services, promote the just allocation of those services and contribute to the education of society regarding language learning and language proficiency.

Principle 8 Language testers shall be mindful of their obligations to the society within which they work, while recognising that those obligations may on occasion conflict with their responsibilities to their test takers and to other stakeholders.

Principle 9 Language testers shall regularly consider the potential effects, both short and long term on all stakeholders of their projects, reserving the right to withhold their professional services on the grounds of conscience.

two decades a growing number of European countries introduced language requirements as part of their immigration and integration policies. Most countries in Europe today have formal language requirements for citizenship, residency and/or entrance to the country, with the level of language proficiency required varying considerably from one country to the next.

Two working groups consolidated their work on ethical issues and inspiring principles in the area of language policy within the migration context:

- The LAMI Special Interest Group (Language Assessment for Migration and Integration) was set up by ALTE in 2002. It aims to represent a platform for language testers in supporting their attempts to ensure test fairness within the migration context.
- The LIAM project (Linguistic Integration of Adult Migrants) was launched by the Council of Europe in 2007 in order to deal with the linguistic challenges imposed by migration flows.

Concrete results of the above-mentioned connection with the Council of Europe welcomed the ALTE-LAMI Booklet *Language tests for access, integration and citizenship: An outline for policy makers* (2016), available in three language versions: English, Finnish and Italian.

It is a position paper that takes into account ethical and technical concerns to ensure that language tests do not discriminate against nor infringe the human rights of migrants. It offers practical guidance in developing responses to some of the major points of reflection raised by the Parliamentary Assembly of the Council of Europe in its Recommendation 2034 (2014).

More recently, the 2018 survey Language policies and language requirements for migrants was organised by the Education Policy Division in close cooperation with the ALTE-LAMI Special Interest Group (SIG) as embedded in the Council of Europe contribution to the United Nations 2030 agenda. The survey was a follow-up to the previous surveys conducted in 2007, 2009 and 2013, with an extended focus on vulnerable groups, such as minors, low-literate migrants and refugees. It had two main purposes in order to allow the formulation of evidence-based policy recommendations:

- map the language and Knowledge of Society requirements in member states, according to different stages of the migrants' journey: prior to entry, residence (temporary and permanent) and naturalisation for third country nationals;
- analyse the development trends in these requirements from 2007 onwards, thus following the previous surveys on this topic, with a specific focus on vulnerable groups (such as minors, refugees and low-literature migrants).

In October 2019, the Council of Europe and ALTE co-organised an intergovernmental conference to present the results of the 2018 survey. Entitled *Achieving Equal Opportunities for All Migrants Through Learning and Assessment: Language and knowledge of society requirements for migrants in Council of Europe member states*, the conference provided insight into the trends in the migration policies in Council of Europe member states, discussed the implications of increasing use of language requirements and suggested concrete measures to be taken.

1.3 The ALTE Code of Practice

In 1994, the Members of ALTE decided that it was essential to adopt a formal Code of Practice which would both define the standards that current and future members would agree to aim to meet in producing their examinations and serve as a statement to consumers of those examinations of what they should expect. The Code of Practice was devised with the principal objectives as stated in its Introduction: in order to establish common levels of proficiency, tests must be comparable in terms of quality as well as level, and common standards need, therefore, to be applied in their production. The Code of Practice sets out these standards and states the responsibilities of both producers and users of language examinations.

THE ALTE CODE OF PRACTICE

As providers of language examinations, the Members of ALTE adopt a Code of Practice in order to make explicit the standards they aim to meet, and to acknowledge the obligations under which they operate. In formulating and adhering to a Code of Practice, it is necessary to distinguish between the various roles of those who have an interest in the issue of setting and maintaining standards in language examinations. These are: **examination developers, examination users** and **examination takers**.

Examination developers are people who actually construct and administer examinations as well as those who set policies for particular testing programmes.

Examination users may select examinations, commission examination development services or make decisions which affect the educational possibilities and careers of others on the basis of examination results.

Examination takers, or candidates, are those who, either by choice or because they are required to do so by examination users, take examinations.

The roles of examination developers and users may of course overlap, as when a state education agency commissions examination development services, sets policies that control the development process, and makes decisions on the basis of the results. Members of ALTE are primarily concerned with the development and administration of examinations. As such, they have a duty towards examination users and ultimately to examination takers. The decisions made by examination users have a direct effect on examination takers or candidates; for that reason, the obligations of examination users are also dealt with in this Code of Practice.

Members of ALTE undertake to safeguard the rights of examination takers by striving to meet the standards of a Code of Practice in four areas:

- Developing Examinations
- Interpreting Examination Results
- Striving for Fairness
- Informing Examination Takers

The Code of Practice is divided into two parts. Part One focuses on the responsibilities of ALTE Members and Part Two on the responsibilities of examination users.

PART 1 – RESPONSIBILITIES OF ALTE MEMBERS

Developing Examinations

Members of ALTE undertake to provide the information that examination users and takers need in order to select appropriate examinations.

In practice, this means that Members of ALTE will guarantee to do the following, for their examinations:

- Define what each examination assesses and what it should be used for.
- Describe the population(s) for which it is appropriate.
- Explain relevant measurement concepts as necessary for clarity at the level of detail that is appropriate for the intended audience(s).
- Describe the process of examination development.
- Explain how the content and skills to be tested are selected.
- Provide either representative samples or complete copies of examination tasks, instructions, answer sheets, manuals and reports of results to users.
- Describe the procedures used to ensure the appropriateness of each examination for the groups of different ethnic or linguistic backgrounds who are likely to be tested.
- Identify and publish the conditions and skills needed to administer each examination.

Interpreting Examination Results

Members of ALTE undertake to help examination users and takers interpret results correctly.

In practice, this means that Members of ALTE will guarantee to do the following:

- Provide prompt and easily understood reports of examination results that describe candidate performance and profiles clearly and accurately.
- Describe the procedures used to establish pass marks and/or grades.
- If no pass mark is set, then provide information that will help users follow reasonable procedures for setting pass marks when it is appropriate to do so.
- Warn users to avoid specific, reasonably anticipated misuses of examination results.

Striving for Fairness

Members of ALTE undertake to make their examinations as fair as possible for candidates of different backgrounds (e.g. gender, age, ethnic origin, special needs, etc.).

In practice, this means that Members of ALTE will guarantee to do the following:

- Review and revise examination tasks and related materials to avoid potentially insensitive content or language.
- Enact procedures that help to ensure that differences in performance are related primarily to the skills under assessment rather than to irrelevant factors such as gender, age or ethnic origin.
- When feasible, provide appropriate accommodation or administration procedures for candidates with special needs

Informing Examination Takers

Members of ALTE undertake to provide examination users and takers with the information described below.

In practice, this means that Members of ALTE will guarantee to do the following:

- Provide examination users and takers with information to help them judge whether a particular examination should be taken, or if an available examination at a higher or lower level should be used.
- Provide candidates with the information they need in order to be familiar with the coverage of the examination, the types of task formats, the rubrics and other instructions and appropriate examination-taking strategies.
- Strive to make such information equally available to all candidates.
- Provide information about the rights which candidates may or may not have to obtain copies of papers and completed answer sheets, to retake papers, have papers re-marked or results checked.
- Provide information about how long results will be kept on file and indicate to whom and under what circumstances examination results will or will not be released.

PART 2 - RESPONSIBILITIES OF EXAMINATION USERS

Examination users are in a position to get information about examinations from examination developers, and a Code of Practice for them concerns the appropriate use of this information. Like examination developers, they have a duty towards candidates, and are under an obligation to set and maintain high standards of fair behaviour. These responsibilities are described below under the following four headings: Selecting Appropriate Examinations, Interpreting Examination Results, Striving for Fairness, Informing Examination Takers.

Selecting Appropriate Examinations

Examination users should select examinations that meet the purpose for which they are to be used and that are appropriate for the intended candidate populations.

Interpreting Examination Results

Examination users should interpret scores correctly.

Striving for Fairness

Examination users should select examinations that have been developed in ways that attempt to make them as fair as possible for candidates of different backgrounds (e.g. gender, ethnic origin, special needs, etc.).

Informing Examination Takers

In cases where the examination user has direct communication with candidates, they should regard themselves as having many of the obligations set out for Members of ALTE under the section in Part One entitled Informing Examination Takers.

(Acknowledgement is made to *The Code of Fair Testing Practices in Education* produced by the Washington D.C. Joint Committee on Testing Practices in 1988 – the latest edition is available here.)

2. Principles of good practice and validation in language assessment

2.1 Achieving good practice

Examinations affect not only individuals, but institutions and society as a whole, as seen previously. The three broad categories of stakeholder identified in the 1994 Code of Practice do not represent the full range of participants in the examination processes. The individuals who are affected by the exams include the takers and their sponsors (students, parents, teachers, job applicants, migrants, employees etc.). In addition there are the professionals and academics involved in the process of designing, writing, administering, marking and validating the exams themselves (i.e. teachers, item writers, consultants, examiners, school owners, test centre administrators, supervisors, textbook writers etc.).
The institutions affected may include schools, universities and colleges, government agencies, publishers, businesses and industry. Individuals and institutions benefit when testing helps them achieve their goals, and society benefits when the achievement of these goals contributes to the general good.

The wide range of stakeholders who can be considered participants in the examination processes includes those who contribute to the production and administration of examinations and those who make use of the test scores and certificates. ALTE Members must be prepared to monitor the views and attitudes of this constituency and to review/change what they do in light of the way these stakeholders use the exams and what they think about them.

The appropriate involvement of these stakeholder groups in examination processes is an important principle in achieving good practice, which also depends on two fundamental principles:

a) the rational planning and management of resources relating to the development, administration and validation of examinations;
b) the collection, storage and use of data/information about all aspects of the examining process.

Failure to capture adequate data means that evidence of standards being reached and maintained cannot be provided (e.g. regarding validity, reliability, impact and practicality).

Good practice and thus high-quality examinations can only be achieved if appropriate procedures are implemented for managing all aspects of the examination process, taking into account new formats and approaches relating to the assessment of language ability. It is therefore necessary to adopt a rational approach (that incorporates the notion of iterative cycles) to examination development, administration and validation. The first stage of this approach must involve planning, including a detailed situational analysis (i.e. a feasibility study) which looks at the perceived need for a new examination within a given educational context, and the most effective way of delivering it. The aim is to identify the considerations and constraints which will be relevant to the examination development project and which will determine how examination usefulness will be achieved.

Whenever it is decided that a new examination development should go ahead, there should be agreed procedures which address at least the following areas:

- the management structure for the development project;
- a clear and integrated assignment of roles and responsibilities;
- a means of monitoring progress in terms of development schedules and resources;
- a methodology for managing the examination production process when the examination becomes operational (i.e. item writing, vetting, moderation, pre-testing, item banking, question paper construction).

Once an examination becomes operational, information must be collected regarding the production of materials and administration of the test in order to judge whether the procedures are meeting expectations regarding aspects of practicality such as cost and efficiency. The cyclical processes which follow the initial planning involve ongoing monitoring of the examination development itself and the subsequent live administrations of the examination. Careful record keeping and data collection is required to monitor all activities. The techniques of monitoring include all kinds of records which serve to establish a documented history of the examination; these in turn serve for both formative and summative evaluation. The data which is collected can be both 'hard data' (empirically collected facts and figures) and 'soft data' (feelings, impressions, attitudes, etc.).

A key aspect of this approach is that validation is an integral part of the process. In order for this to occur, the procedures which are implemented for the ongoing production and administration of the examination should be designed so that adequate data can be collected, stored and retrieved as required.

As a principle, it will be the overall usefulness of an examination that must be maximised, in order to generate accurate and meaningful feedback. This means that it is

inevitable that evidence collected regarding one aspect of an examination will be relevant to the others. For example, data collected on examination reliability will not only be used in a narrow way to address the question of reliability, but will also be used to address questions of validity, fairness, and impact (as defined in the next section). All of this provides validation, based on sound argument, to ensure the quality of the examination in question.

2.2 The concept of usefulness in examinations

The principles of good practice proposed here are aimed at ensuring that examinations offered by ALTE Members can be shown to meet explicit criteria in terms of the following qualities of examination services:

- Content validity
- Construct validity
- Reliability
- Criterion-related evidence
- Fairness
- Quality of service
- Practicality
- Impact

These qualities must run throughout the preparation and deployment of the test task(s), the observation of the performance, and the resulting feedback which is presented to the test taker and to other users, stakeholders and society as a whole. The usefulness of these qualities will be evidenced through argument-based validation.

By addressing these aspects of their examination services in a principled way, the Members of ALTE ensure that their commitments are met and sufficiently high standards can be maintained.

Not surprisingly the qualities of content, construct and criterion-related validity, as well as of reliability, have been discussed extensively in the literature on measurement and language testing. For example, the AERA/APA/NCME *Standards* (both the 1999 and 2014 editions) provide extensive discussions and Bachman (1991) dedicates a chapter to each. The other qualities have always been important considerations for examination developers but have only recently emerged in the literature of language testing. It is now broadly recognised that the individual examination qualities cannot be evaluated independently and that the relative importance of the qualities must be determined in order to maximise the overall usefulness of the examination (see for example in Bachman and Palmer 1996).

The concept of examination usefulness requires that, for any specific assessment situation, an appropriate balance must be

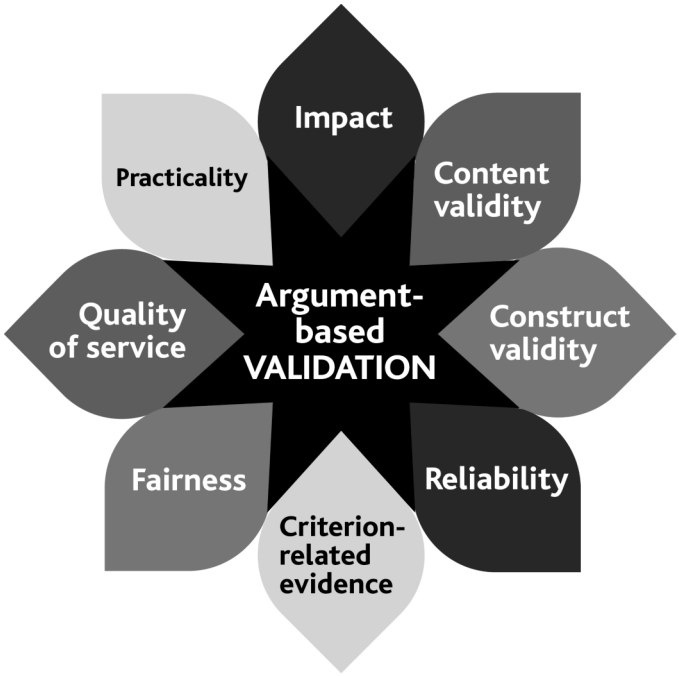


Figure 3: Aspects of an argument-based validation of good practice in language examining

achieved between these qualities. It is recognised that Members of ALTE should be held accountable for all matters related to use of their examinations; this involves providing a high-quality service to the users of their examinations which meets the principles of good practice as outlined in this document.

All examinations are context-specific as they meet the needs of the plurilingual and pluricontextual learner and this means that practical considerations and constraints must be taken into account regarding examination development and examination administration so that the appropriate balance between the examination qualities is achieved for any given situation (e.g. educational context, group of examination takers and examination purpose). The relative importance of the qualities must be determined in order to maximise the overall usefulness of the examination. Successful examinations cannot be developed, however, without due consideration being given to all qualities.

In considering the context in which an examination is to be developed and used, it is necessary to take into account the specific considerations and constraints which characterise that situation. These will not be the same for all ALTE examinations and will determine whether an examination is feasible and can be produced and administered with the resources available. With regard to resources, this applies to the resources which are available internally to the ALTE institutions and also to the resources which are available in the contexts where the examinations will be administered. In particular, costs for both development and administration must be controlled and managed.

2.3 Examination qualities

As Figure 3 shows, ALTE Members should ensure good practice in relation to the following qualities of their examination services: content validity, construct validity, reliability, criterion-related evidence, fairness, quality of service, practicality, and impact; and should be prepared to demonstrate the evidence for these – thus providing a useful test to their audiences.

2.3.1 Content validity

Validity as a whole concerns the appropriateness and meaningfulness of an examination in a specific educational context and the specific inferences made from examination results. 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 candidate's true level of language ability. Although validity can be seen as a unitary concept (cf. Messick's chapter on validity in Linn 1989, AERA, APA and NCME *Standards* 2014:14), 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. The 2014 edition of the *Standards* highlights the need to refer to types of validity evidence rather than to distinct types of validity:

> *Statements about validity should refer to particular interpretations and consequent uses. It is incorrect to use the unqualified phrase 'the validity of the test'. No test permits interpretations that are valid for all purposes or in all situations. Each recommended interpretation for a given use requires validation. The test developer should specify in clear language the population for which the test is intended, the construct it is intended to measure, the contexts in which test scores are to be employed, and the processes by which the test is to be administered and scored. The overarching standard for validity is then clear articulation of each intended test score interpretation for a specified use and provision of appropriate evidence in support of each intended interpretation.*
>
> *(AERA, APA and NCME Standards 2014:23)*

Content-related evidence is **the extent to which the test covers the full range of knowledge and skills relevant and useful to real-world situations and authentic language use**. This is important as the test tasks need to assess the performances of the real-life tasks that the plurilingual, pluricontextual language learner finds, or will find, him- or herself in.

Content-related validation investigates the degree to which the sample of items, tasks, or questions on an examination are representative of a defined domain of content. It is concerned with both relevance and coverage.

A wide range of methods for testing the representativeness of the sample are available; major characteristics of the domain can be specified through a model (e.g. the Waystage and Threshold specifications) and experts in the field can be asked to assign examination items to the categories defined by these characteristics; in this way the representativeness of the content can be judged.

The specification of the domain of content that an examination is intended to represent is very important; the degree to which the format of items or tasks in an examination are representative of the domain is crucial; and the involvement of stakeholder groups and relevant experts is a key element in the process of test development. Often systematic observations of behaviour in the 'real world' can be used to identify distinctive features or characteristics of the criterial situation (cf. Bachman and Palmer's Target Language Use – TLU – domain, 1996:44–45). These observations may be combined with expert judgements to build up a representative sample of the content domain. A concern for the authenticity of test content and tasks, and the relationship between the 'input' and the expected response or 'output', is an important feature of content validation. The authenticity of the tasks and materials in the examinations, giving due consideration to the diversity of the test takers and test users, can be considered a major strength of the approach to assessment they adopt. The examination content must be designed to provide sufficient evidence of the underlying abilities (i.e. construct) through the way the test taker responds to this input. The responses to the test input (tasks, items, etc.) occur as a result of an interaction between the test taker and the test content. The authenticity of test content and the authenticity of the candidate's interaction with that content are important considerations for the examination developer in achieving high validity. (See Widdowson 1978, 1983 on situational and interactional authenticity and Bachman and Palmer 1996 for the application of these concepts to language tests). This is why it is important to recognise the plurilingual, pluricultural language learner at the centre of assessment, and to build an argument of content validity around the situations and contexts that the language learner needs to prove communicative competence in.

In summary, content-related validation is linked to examination construction, as well as to establishing evidence of validity after the examination has been through the developmental phase and is considered 'live'.

2.3.2 Construct validity

Construct-related evidence is the **extent to which the test results conform to the model of communicative language ability underlying the test**.

The focus in construct validation is primarily on the examination score or grade as a measure of a trait or 'construct'. The examination developer defines traits of ability for the purpose of measurement and it is these definitions which are the constructs. A model of communicative language ability represents a construct in the context of language testing (cf. Bachman 1991, Canale 1983, Canale and Swain 1980). By focusing tests on the language learner's social world, as well as cognitive abilities, Weir (2005) developed the socio-cognitive framework which further serves to provide construct validity. It is in this way the plurilingual, pluricontextual language learner is now served by language assessment.

The process of compiling construct-related evidence for examination validity starts with examination development and continues when the examination 'goes live' and is used under operational conditions.

Validating inferences about a construct requires paying great attention to many aspects of measurement, such as examination format, administration conditions, or level of ability, which may affect examination meaning and interpretation.

Construct validity is seen by many as the 'unifying concept' within test validation that incorporates content and criterion considerations (Messick 1989). As a process, construct validation seeks evidence from a variety of sources in order to provide information on construct interpretation. The choice of which approach to use in gathering evidence for the interpretation of constructs depends on the particular validation problem and the importance of the role of given constructs within the investigation. In the literature, a wide range of statistical techniques have been used, largely based on correlations and often using experimental designs to collect data.

2.3.3 Reliability

Reliability is a key concept in any form of measurement and contributes to overall validity.

In language assessment, reliability concerns **the extent to which test results and feedback are precise, stable, consistent, and free from errors of measurement**. In other words it concerns the degree to which examination marks can be depended on for making decisions about the candidate. Estimates of reliability should not only consider relevant sources of error, but the types of decision which are likely to be based on examination marks.

For a wide variety of reasons individuals may score differently on two forms of an examination which are intended to be parallel; when these differences cannot be accounted for, they are called **errors of measurement**. Measurement errors reduce reliability (and thus the generalisability) of marks obtained for an individual from a single measurement.

Reliability is generally estimated and reported in terms of **reliability coefficients**. Since this is a generic term, the information about error it conveys varies with the specific estimation method used, and since not all sources of error will be relevant to every examination, it is the responsibility of the examination developer to decide on appropriate forms of reporting error variance. This may involve reporting standard errors of measurement, confidence intervals, dependability indices etc.

Within language testing, much of the literature on computing the reliability of language tests has been based on work in educational and psychological testing more generally, e.g. the AERA, APA and NCME *Standards* between 1954 and 1985. In the 1999 volume of the *Standards* the revised chapter on Reliability and Errors of Measurement (Part 1, Section 2) identified three broad categories of reliability which have traditionally been recognised in the field:

a) alternate-form coefficients (derived from the administration of parallel forms in independent sessions);

b) test-retest coefficients;

c) the use of internal consistency coefficients, such as Cronbach's alpha or KR20 to estimate the reliability of objective tests is common (e.g. for multiple-choice reading or listening tests). The fact that these coefficients are relatively easy to calculate mean that other, perhaps more appropriate estimates, are not used as commonly (e.g. test-retest estimates are less often reported because adequate data is difficult to obtain under operational conditions).

The overarching standard for reliability/precision offered in the newest edition of the *Standards* (2014) stresses that the forms of evidence for reliability/precision, such as reliability or generalisability coefficient, item response theory (IRT), conditional standard error, and index of decision consistency should be appropriate for the intended uses of the scores, the population involved, and the psychometric models used to derive the scores (AERA, APA and NCME *Standards* 2014: 42).

For tests of speaking and writing the *Standards* make it clear that when the scoring of a test involves judgement by examiners or raters, it is important to consider reliability in terms of the accuracy and consistency of the ratings which are made. The tests of speaking and writing found in many of the exams offered by ALTE Members fall into this category because the assessments are made by examiners.

The reliability of subjective assessments (using examiners) is usually estimated using correlations, e.g. intra- and inter-rater correlations.

In providing evidence of reliability for the interpretation of each intended score use, good practice should involve at least the following:

a) Serious efforts to identify and quantify major sources of measurement error, including:
   - the degree of reliability expected between pairs of marks in particular contexts (e.g. marks achieved by a candidate on two different tasks which are intended to be of equivalent difficulty);
   - the generalisability of results across tasks and items, different forms of the same exam, examiners, different administrations, etc.

b) An assurance that examination marks, including sub-scores and combinations of marks, are sufficiently reliable for their intended use.

c) Provision of information on reliabilities, standard errors of measurement, or other equivalent information so that examination users can also judge whether reported examination marks are sufficiently reliable for their intended use.

d) Provision of information for examination users about sources of variation and other sources of error considered significant for score interpretation.

e) Estimates of the reliability or consistency of reported examination marks by methods that are appropriate to the nature and intended use of the examination marks and that take into account sources of variance considered significant for score interpretation.

f) Documentation of the reliability analysis, including:
   - a description of the methods used to assess the reliability or consistency of the examination marks and the rationale for using them, the major sources of variance accounted for in the reliability analysis and the formula used and/or appropriate references;
   - a reliability coefficient, an overall error of measurement, an index of classification consistency, or other equivalent information about the consistency of examination marks;
   - standard errors of measurement or other measures of mark consistency for mark regions within which decisions about individuals are made on the basis of examination marks;
   - the degree of agreement between independent markings when judgemental processes are used;
   - correlations among reported sub-scores within the same examination or the marks within an examination battery.

g) Descriptions of the conditions under which the reliability estimates were obtained, including:
   - a description of the population involved, e.g. demographic information, education level, employment status;
   - a description of the selection procedure for, and the appropriateness of, the analysis sample, including the number of observations, means, and standard deviations for the analysis samples and any group for which reliability is established;
   - when marks are based on judgements, the basis for marking, including selecting and training markers, and the procedures for allocating papers to examiners and adjusting discrepancies;
   - the time intervals between examinations, the rationale for the time interval and the order in which the forms were administered if alternate-form or test-retest methods were used.

2.3.4 Criterion-related evidence

Criterion-related evidence (predictive and concurrent validity) **is the extent to which test scores correlate with a recognised external criterion which measures the same area of knowledge or ability** (e.g. with reference to a system of levels such as the CEFR).

Criterion-related validation aims at demonstrating that test scores or examination grades are systematically related to an external criterion or criteria (e.g. another indicator of the ability tested). It is the criterion, therefore, that is of central interest. This criterion may be defined in different ways; for example, by group membership, by performance on another examination of the same ability, or by success in performing a real–world task involving the same ability.

When ALTE was established there were no recognised international frameworks, and language certification was variable in terms of the levels to which it referred. However, examination developers and other users were beginning to become aware that they needed a mechanism to understand levels and what they meant, and how exams in different languages related to each other both in relation to content and level.

It was in fact in this context that ALTE was formed and its objectives articulated, with the first steps in collaboration being to produce descriptions of the Members' examinations and to place them in a grid against a common proficiency scale. Systematic comparisons between exams in different languages thought to be at corresponding levels of the framework were carried out using Question Paper Content Checklists, which were translated into a number of languages. Validation of the grids through experimental work was then carried out through the 'Critical Levels Project' which resulted in 'Can Do' statements and paved the way for the development of the CEFR.

The Council of Europe's second Rüschlikon Conference in November 1991 was very closely linked to early developments in ALTE. Rüschlikon launched the CEFR for language teaching, learning and assessment. Much work took place through the 1990s, a lot of it in parallel with developments taking place in ALTE, though ALTE's focus was largely in the area of assessment and less so in relation to the learning and teaching of languages at that time. However, the approach taken to creating a 'Can Do'-oriented level system based on calibrated statements using Rasch analysis was very similar, although the ALTE approach also linked the 'Can Do' statements to performance in exams across a range of different languages.

The original ALTE Framework was therefore complementary to the framework in the CEFR. The ALTE Framework was an operationalisation of the CEFR for examination purposes.

Table 1: ALTE levels and corresponding CEFR levels

ALTE level	*CEFR level*
Level 5 (Good, later Mastery)	C2
Level 4 (Competent, later Proficient)	C1
Level 3 (Independent, later Vantage)	B2
Level 2 (Threshold)	B1
Level 1 (Waystage)	A2
Breakthrough	A1

In the literature, there are two specific kinds of criterion-related evidence which are discussed – concurrent and predictive.

Providing evidence for concurrent validity involves obtaining information on the accuracy with which examination data can be used to estimate or predict criterion-related behaviour. The most common information is based on correlations between various measures which are made concurrently (e.g. to show the relationship between scores on two different tests of the same ability). In the case of performance tests, qualitative comparisons can be made between the criterion norms and samples of the output from the test (e.g. in writing or speaking).

Evidencing predictive validity serves a similar purpose but obtains predictive information in relation to the future, such as future examination results, performance in higher education or performance in a future job. Evidence of this kind of validity is particularly important where the examination or test results are used for screening or placement purposes.

2.3.5 Fairness

Fairness is the third fundamental consideration accompanying validity and reliability in the 2014 edition of the AERA, APA and NCME Standards. In previous versions fairness was presented in various chapters, related to the assessment of test takers with disabilities and diverse linguistic and cultural backgrounds. In the 2014 edition, these issues are presented in a separate chapter (Chapter 3), offering a set of 20 standards that emphasise fairness as an overriding concern that may interfere with the validity of test score interpretation. These standards describe the different aspects of fairness, such as the fair and equitable treatment of all test takers during the testing process, lack of measurement bias, access to the construct measured, and fairness as validity of individual test score interpretation for the intended use. The overarching standard for fairness presented as the guiding principle of the chapter requires that 'all steps in the testing process, including test design, validation, development, administration and scoring procedures be designed in such a manner as to minimise construct-irrelevant variance and to promote valid score interpretations for the intended uses for all examinees in the intended population' (AERA, APA and NCME *Standards* 2014:63). Challenges which test takers may face in demonstrating their ability may arise due to their disability, cultural, linguistic or educational background, socioeconomic status, age, or a combination of these and other factors. With standardisation remaining fundamental for assuring the same opportunity to all examinees to relate to the construct measured by the test, flexibility is sometimes needed to provide equivalent opportunities for some test takers. For example, a Braille test form, a screen reader, or a large-print examination paper may be provided to enable visually impaired examinees to have more equitable access to test content. Any test adaptation must be, however, carefully considered in order not to obstruct the intended construct of a test. A fair test should not advantage or disadvantage some test takers because of any construct-irrelevant characteristics. A fair test offers the same construct to all test takers and its score has the same meaning for all in the intended population.

Good practice in addressing issues of fairness in testing should take into account both the diversity of test takers and test users, and the different testing formats involved. Good practice will also include the following:

- ensuring fair and equitable treatment of all test takers during the testing process;
- providing equitable access to the construct measured;
- monitoring and minimising measurement bias;
- ensuring validity of score interpretation for the intended use of individually adapted tests.

2.3.6 Quality of service

Quality of service concerns **the examination developer's ability to meet specific commitments to the examination takers and users**. This includes the provision of secure examination materials, the confidentiality of examination data and results, and procedures to handle enquiries about results and appeals procedures.

Good practice in achieving high quality of service should include the following:

Quality of service in research:

- Making available the results of research, and seeking peer review of such activities.
- Asking for information about individuals and institutions only when it is potentially useful to them by way of furthering the research on products, and thereby improving them. The purpose of gathering such information should be made clear to everyone concerned.

Quality of service in delivery:

- Making realistic delivery commitments and subsequently making every effort to meet these commitments.
- Using adequate quality controls to ensure that any products and services offered by ALTE Members are accurate and delivered within the time spans promised.

Quality of service in content:

- Reviewing and revising examination questions and related materials in order to avoid potentially insensitive or inappropriate content and language.
- When feasible, making appropriately modified forms of examinations or administrative procedures available for candidates with special requirements or needs.

Quality of service in information:

- Providing candidates with the information they need in order to be familiar with the coverage of the examination, the types of question formats, rubrics, appropriate test-taking strategies and how the test will be marked.
- Striving to make such information equally available to all candidates.
- Telling candidates how long their results will be kept on file and indicating to whom and under what circumstances examination results will or will not be released.

Quality of service in data:

- Protecting the confidentiality of all data (raw or processed) held by ALTE Members on any institutions or individuals, and encouraging any group or institution to or from which data is transferred to adopt the same policy.

Quality of service in nonconformities:

- Accepting the responsibility for informing those negatively affected if, subsequent to its release, information is found to be inaccurate.
- Informing those negatively affected if there is likely to be a substantial departure from scheduled commitments.
- Describing the procedures that the candidates may use to register a complaint or an appeal and have their problems resolved.

2.3.7 Practicality

In order for a test to be useful, it must be practical to implement. This factor is involved in many, if not all of the qualities described in this chapter, as it requires a balance between the qualities. Whether a test is practical depends on what is needed to develop and deploy it, thinking about the context in which the test will be used.

The practicality of any examination involves two factors:

a) the resources that are required to produce an operational examination that has the appropriate balance of qualities (e.g. of validity, reliability and impact) for the context in which the examination will be used;

b) the resources that are available.

A practical examination is one that **does not place an unreasonable demand on available resources**. If available resources are exceeded, then the examination is not practical. In this case, the examination developer must either modify the design of the examination or make a case for an increase or reallocation of resources. If practical constraints make the second option impossible, the first option must be chosen.

Before the examination development can proceed, it must be established whether or not the examination will still be useful if the changes to the specifications are implemented. If this cannot be established, the examination development should not proceed.

Good practice in ensuring practicality should include the following:

- good practice can only be achieved if appropriate procedures are implemented for managing all aspects of the examination process, including the development, administration and validation of the examinations;
- the development of practical examinations requires that an explicit model of test development be adopted – see for example the *Manual for Language Test Development and Examining*;
- whenever a new or revised examination is to be developed, there should be procedures in place to address the management structure for the development project with a clear and integrated assignment of roles and responsibilities;
- there needs to be a means of monitoring progress in terms of development schedules and resources, and a methodology for managing the examination production process when the examination becomes operational (item writing, vetting, moderation, pre-testing, item banking, question paper construction).

The process of development should begin with a feasibility study dealing with at least the following:

- the purpose of the new examination;
- the level of difficulty for the intended examination takers (e.g. in relation to the CEFR);
- external factors – the market place, the competition provided by existing exams of a similar kind, societal demands or requirements (e.g. from parents, Ministries of Education, etc.);
- intrinsic factors: theories related to the examination constructs and content, advances in technology, fixed institutional parameters;
- the predicted relevance and acceptability of the new examination to intended takers and users;
- the cost of the new examination to the taker;
- the resources available for:
  - development
  - administration
  - reporting of results
  - replication (future administrations of the examination).

The determination of examination usefulness is both cyclical and iterative; considerations of practicality affect decisions at all phases of the examination development process. When operational considerations are taken into account, it is necessary to consider to what extent it is possible to achieve this balance with the resources that are available (e.g. in terms of people, equipment, time and money).

2.3.8 Impact

The literature on impact is borne out of models of washback (Alderson and Wall 1993, Green 2003, etc.) which then evolved into impact more broadly (e.g. Cheng 2005, Green 2007, Wall 2005). Building on Milanovic and Saville (1996), Saville (2009) proposed a model which involved 'impact by design', meaning designing tests which have the potential for positive impacts, 'anticipating the possible consequences of a given policy "before the event"'.

It is recognised that, as providers of examinations, Members of ALTE have a major impact on educational processes and on society in general and consideration as to how to achieve positive impact must not be an afterthought. The assessments taken by the plurilingual, pluricontextual language learner provide feedback, and this and the testing process as a whole has impact which operates on at least two levels:

a) a macro level in terms of general educational processes;
b) a micro level in terms of the individuals (stakeholders) who are affected by examination results.

One area of general impact concerns the role of ALTE in promoting the public understanding of assessment and related pedagogical issues within Europe and worldwide. This can be achieved by providing public information, research and advisory services. The aim should be to achieve greater understanding of the purposes and procedures of testing and the proper uses of examination information (results, grades, etc.).

In terms of impact on individuals, it is necessary to establish that the examination is fair and not biased.

Positive impact on teaching and learning is an important aspect of impact which operates on both levels (macro and micro). It is in this context that the notions of 'face validity' (or test appeal) and washback are considered. It is important to be able to investigate the educational impact that examinations have within the contexts in which they are used. As a point of principle, examination developers must operate with the aim that their examinations will not have a negative impact and, as far as possible, strive to achieve positive impact.

In providing evidence of impact, good practice should involve at least the following:

a) the development and presentation of examination specifications and detailed syllabus designs;
b) provision of professional support programmes for institutions and individual teachers/students who use the examinations.

Positive educational impact can also be achieved through the following practices:

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

Procedures also need to be put into place when an examination becomes operational in order to collect information which allows impact to be estimated.

This should involve collecting data on the following:

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

This information should be made available within the ALTE organisations, for example in the form of written reports, and suitable versions of such reports should be made available to the other stakeholders.

From the evidence collected, it should be possible to demonstrate that the examination is sufficiently valid and reliable for the context in which it is used. This in itself is a way of ensuring that positive impact is achieved.

2.4 Conclusions

In providing evidence of usefulness, good practice should involve at least the following:

a) A description of the constructs to be measured and the domain of content covered by the examination.
b) Evidence related to the use of examination results, including a description of how the evidence provided is appropriate for the inferences that are drawn and the actions that will result from examination results.
c) A description of the validation procedures used and their results including as appropriate:
   - logical and empirical analyses of processes underlying performance in examinations;
   - the relationship between examination results and other variables, including likely sources of variance not related to the construct;
   - how the examination questions/items were derived and are related to the domain of knowledge or skill appropriate to the intended inferences to be made;

- logical and empirical evidence supporting discriminant validity of sub-scores;
- the number and the qualifications of any experts who made judgements which are pertinent to the validation process;
- procedures used to arrive at judgements, which are pertinent to the validation process;
- the rationale and procedures used in designing the examination specifications (including range of materials surveyed, etc.);
- the rationale and procedures for determining criterion relevance;
- information relating to the interpretation of quantitative evidence.

d) The carrying out of new studies on validity whenever there is a substantial change in the examination, the mode of administration, the characteristics of intended examination takers, or the domain of content to be sampled.

e) The provision of information to examination users to help them interpret validation studies with respect to intended examination results, such as pass/fail decisions, selection or placement.

3. ALTE Quality Management System

3.1 Introduction

Having established the principles and provided some practical tools to help ALTE Members improve their examination systems, the Association addressed the issue of how to put the principles into practice, how improvements can be monitored and whether adequate standards are in fact being met. While most people in ALTE agreed with the principles, it was more difficult to get consensus on how the standards could be set in an equitable way, allowing for the diversity of organisations and testing practices across ALTE as a whole. Van Avermaet, Kuijper and Saville (2004:144) noted the:

> *... differences between the [ALTE] Members with respect to the organisational, linguistic, educational and cultural contexts within which the examinations are being developed and used. Furthermore, within the institutions themselves there are huge differences in knowledge and traditions with respect to statistical and empirical issues such as data gathering, data analysis, equating different examinations, and so forth. In the early discussions of the Working Group, these differences were looked at from the point of view of the different organisational types and the examination systems that are currently in place. They realized that introducing a system of quality control could be very threatening for some members who know for themselves that they do not meet high standards at the moment—particularly when compared with other members of the association. The approach to QMS [Quality Management Systems] that was adopted was therefore designed to lower anxiety and was meant to be a supportive tool. The aim was to allow members (a) to enhance the quality of their examinations in the perspective of fairness for the candidates; (b) to engage in negotiations with their senior management and sponsors in a process of organisational change, where necessary (e.g., to ensure that resources are made available to support ongoing improvements).*

In order to address this problem and to seek consensus, it was decided that the appropriate paradigm for this activity would be that of Quality Management Systems (QMS) (such as that represented by the ISO 9000 series). QMS seek to improve the products and/or services of an organisation in order to meet the requirements of its customers in the most effective way, and they go about doing so in a well-planned and focused manner. In adopting the QMS approach, Members undertook to understand the nature of their organisations better and in so doing to involve their stakeholders in striving for improvements in quality. As Van Avermaet et al (2004:149) explained:

> *The QMS functions as a tool for members to enhance the quality of their examinations in the perspective of fairness for the candidates. It can also function as a tool for discussions and debates among partners about the quality and the aspects of fairness of the different procedures and steps in the running of exams. And finally it can also be a tool to open up discussions and negotiations with the funders of examination providers. It is often increased funding and other aspects of organisational change that will lead to greater possibilities of increasing quality and raising standards.*

In order to provide a tool to raise awareness of those areas where change was necessary, Quality Assurance Checklists were developed to reflect the four aspects of the testing cycle with which all ALTE Members and other test developers are very familiar:

- examination development;
- administration of the examinations;
- processing of the examinations (including the marking, grading and issue of results);
- analysis and post-examination review.

The QMS approach is intended to be a supportive tool to allow Members to:

- enhance the quality of their examinations from the perspective of fairness for the candidates;
- engage in negotiations with their senior management and sponsors in a process of organisational change, (e.g. to ensure that resources are made available to support on-going improvements);
- move from self-evaluation to the possibility of external verification in order to set agreed and acceptable standards.

The QMS underpins the ALTE auditing system which is explained in the next section.

3.2 The ALTE auditing system: monitoring standards – auditing the quality profile

The formal scrutiny of standards is the culmination of a long process of working towards establishing audited 'quality profiles' across ALTE Members' examinations. The aim of the auditing process is to allow ALTE Members 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, based on 17 parameters for establishing Minimum Standards. Ultimately, this is to ensure that the assessment is fair and meets the needs of the stakeholders in appropriate ways. It is important to remember the context in which this work has been carried out and in particular the wide range and diversity of ALTE Members and Affiliates:

- 33 Full Members: these are organisations which include government departments, universities, consortia and examination boards which have a role in assessing their own language;
- 25 languages are currently represented, including many less widely taught languages;
- in addition, there are approximately 60 Institutional Affiliates and over 500 Individual Affiliates with an interest in language education and language assessment.

The description of the development and application of the ALTE auditing system ties in with recent discussions in the language testing literature on the use of argumentation to support claims of validity (see section 2.3.1 and 2.3.2. in this document). The system is also an example of another kind of theory, 'a theory of action' or change processes (see Fullan 1993, 1999). A theory of action, capable of bringing about positive and sustainable change, was needed by ALTE Members in order to raise standards and improve the quality of their examinations.

It was always envisaged that ALTE Member self-evaluation would need to be supplemented by an external 'auditing' system. This was developed and piloted starting in 2005–06 and continuing with the first cycle from 2007–08, seeing audited 'quality profiles' established across a wide range of ALTE Members' examinations. Taking the Code of Practice (see section 1.3) and Quality Assurance Checklists into account, 17 parameters for establishing Minimum Standards (MS) were agreed, with the aim of establishing a Quality Profile for each exam or suite of exams. The Quality Profile is created by building a validity argument which explains how the examination meets the Minimum Standards, and provides adequate evidence to support the claim. The formal external scrutiny of these parameters in the auditing process is intended to ensure that adequate standards are being set and achieved. ALTE Members are required to make a formal, ratified claim that a particular test or suite of tests has a quality profile appropriate to the context and use of the test, bearing in mind the following points:

- Different tests are used in different contexts, by different groups of test users. There is no intention to impose a single set of uniform quality standards across all ALTE Members' exams.
- Members requesting an audit of their quality systems and procedures are invited to build a validity argument that the quality standards within a test or suite of tests are sufficient and appropriate for that test or suite of tests.
- It is the validity argument which is the subject of the audit, rather than the organisation itself (which is often dealt with by other systems of regulation, e.g. ISO 9001, government regulators etc.).
- Each audit considers one test, suite of tests or testing system.
- The audit has both a consultancy and quality control role.
- The audit aims to establish that minimum quality standards are being met in a way that is appropriate to the context of a test, and also to offer recommendations towards best practice where, though quality standards are appropriate, there is still room for improvement.
- If quality standards are not being met, ALTE auditors, who are ALTE Members, will collaborate with the audited organisation to implement an action plan aimed at working towards and ultimately reaching the quality standards.

3.2.1 Description of a validity argument

The argumentation structure is as follows:

- A **claim** is made about each of the 17 parameters; the claims support an argument that the MS are being met for the test in question.
- **Information** is provided to support the claims; this information is provided in the form of explanations.
- A **justification** is also needed to provide legitimacy for this information; this must be based on the relevant language testing theory with reference to the Code of Practice and the PoGP, and may also take into account prior experience and best practice models where appropriate.
- This justification, in turn, needs to be backed up with **appropriate evidence** which has been collected as part of the validation processes. This approach is consistent with Toulmin's (2003) argument structure and Bachman's (2005) application of his work to language testing. In Toulmin, the justification is known as a **warrant**, and the evidence is known as the **backing**. Bachman (2005) suggests that the test developer must be prepared to deal with **rebuttals** (alternative explanations and counter claims) and to provide additional evidence to reject them (rebuttal data).

3.2.2 Building a validity argument – ALTE Minimum Standards for establishing quality profiles in language testing

The explanation of a validity argument above illustrates how the ways in which language tests can achieve quality (practicality, reliability, quality of service, fairness, content validity, construct validity, criterion-related validity and impact) are borne out in practice by minimum standards, which are at the core of the ALTE auditing system.

ALTE has established a set of common standards for its Members' exams (2007), which cover all stages of the language testing process: test development; task and item writing; test administration; marking and grading; reporting of test results; test analysis; and reporting of findings.

ALTE MINIMUM STANDARDS

TEST CONSTRUCTION

1. You can describe the purpose and context of use of the examination, and the population for which the examination is appropriate.
2. The examination is based on a theoretical construct, e.g. on a model of communicative competence.
3. You provide criteria for selection and training of constructors, expert judges and consultants in test development and construction.
4. Parallel examinations are comparable across different administrations in terms of content, stability, consistency and grade boundaries.
5. If you make a claim that the examination is linked to an external reference system (e.g. the CEFR), then you can provide evidence of alignment to this system.

ADMINISTRATION & LOGISTICS

6. All centres are selected to administer your examination according to clear, transparent, established procedures, and have access to regulations about how to do so.
7. Examination papers are delivered in excellent condition and by secure means either physically or electronically to the authorised examination centres, your examination administration system provides for secure and traceable handling of all examination documents, and confidentiality of all system procedures can be guaranteed.
8. The examination administration system has appropriate support systems (e.g. phone hotline, web services etc.).
9. You adequately protect the security and confidentiality of results and certificates, and data relating to them, in line with current data protection legislation, and candidates are informed of their rights to access this data.
10. The examination system provides support for candidates with special needs.

MARKING & GRADING

11. Marking is sufficiently accurate and reliable for purpose and type of examination.
12. You can document and explain how reliability is estimated for rating, and how data regarding achievement of raters of writing and speaking performances is collected and analysed.

TEST ANALYSIS

13. You collect and analyse data on an adequate and representative sample of candidates 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.
14. 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 candidates and analysed.

COMMUNICATION WITH STAKEHOLDERS

15. The examination administration system communicates the results of the examinations to candidates and to examination centres (e.g. schools) promptly and clearly.
16. 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.
17. You provide suitable information to stakeholders to help them interpret results and use them appropriately.

When preparing for an audit, the auditee completes an ALTE Validity Argument of the Auditee form, which enables the auditee to build a case or argument for their examination(s) by providing information for each Minimum Standard (MS):

1. A description of what is done to meet the MS.
2. A description of why doing this adequately meets the MS.
3. Evidence of what is done and that it is adequate.

The 17 Minimum Standards are available in a range of languages:

- Български
- Català
- Cymraeg
- Čeština
- Deutsch
- Eesti
- English
- Español
- Euskara
- Français
- Gaeilge
- Galego
- Italiano
- Lietuvių
- Magyar
- Nederlands
- Norsk (bokmål og nynorsk)
- Polski
- Português
- Русский
- Slovenščina
- Suomi
- Svenska

3.3 Overview of an ALTE audit

In 2007, a manual named the *Procedures for Auditing* was developed. This document, which sets out the practicalities of the process of an ALTE audit, was updated in 2017. Auditors are recruited, trained and appointed from within the ALTE membership and are required to attend training once a year in order to continue as an auditor. The ALTE membership as a whole is the arbiter of decisions arising from the auditing process; this takes place through the Council of Members as a whole and in particular through the smaller, elected Standing Committee which has delegated responsibility to oversee the

auditing process. The following is a brief outline of the audit process. For full details see the *Procedures for Auditing (PfA).*

- Before applying to go through an audit the **auditee** must attend an audit **orientation session**.
- Once appointed, the auditor liaises with the auditee, then reviews the auditee's validity argument, either face-to-face or remotely, and submits the final audit report to the Standing Committee.
- Two super readers, who are elected Standing Committee members, scrutinise the claims and the evidence; they can challenge whether the claims adequately meet the MS, and if necessary can ask for additional information to be provided.
- An audit remains 'ongoing' or 'in progress' until such points have been clarified or alternative procedures have been put in place which are deemed acceptable.
- Each MS is categorised as one of the following: **Good practice** (GP), **Satisfactory with recommendations for improvement** (RFI) or **In need of improvement** (INI).
- GP and RFI mean the MS has been met. INI means the Minimum Standard has not been met.
- The outcome for the whole audit is either **Resolved** or **Unresolved**.
- When the outcome of an audit are **Unresolved**, the Standing Committee request that an **Action Plan** should be implemented.

3.4 Continual development of the auditing system

After each audit the auditor has to write an **audit report**, which is first discussed with the auditee and then sent to the Standing Committee to be discussed and ratified. In the early years of the auditing system there were significant discrepancies between the reports despite the fact that all the auditors followed the guidelines as described in the *Procedures for Auditing*. This probably had to do with the following issues:

- differences in background of the auditors, leading to a different focus of attention;
- a need for greater elaboration of the core elements within each minimum parameter to achieve better standardisation of the pre-audit, the audit and the reporting;
- a need for more clarity in, and agreement on, which core elements have to be met in order to meet the MS for each parameter.

Comparisons of the information in the different audit reports, and the way different auditors came to their judgements, prompted a review of the ALTE Validity Argument of the Auditee form to provide a more accurate description of the core elements of each minimum parameter. The review was carried out by the QMS special interest group.

Working in this way it is possible to make audits more comparable, transparent and less dependent on individual interpretations of the different auditees and auditors.

The completed audits have also provided a useful 'snap shot' of the state of affairs across the examinations of ALTE Members. This information functions as an input for further training and for organising well-targeted workshops to help improve examinations. The auditor training itself has developed based on the auditing experiences described above.

In conclusion, the QMS and the auditing procedures are a dynamic process which enables the continual monitoring of standards and are useful for:

- clarifying the quality demands of examinations in relation to their functions and purposes;
- providing ALTE Members with valuable information of the state of affairs in the examinations in their frameworks;
- providing ALTE Members with clear guidelines for improving their examinations;
- setting priorities for training, workshops, consultancy and support;
- accounting for the validity of the examinations to stakeholders;
- improving the QMS and auditing systems.

3.5 The ALTE Q-mark

The ALTE Q-mark is a **quality indicator** which Member organisations can use to show that their exams have passed a rigorous ALTE audit and meet the core requirements of all 17 ALTE's MS. The Q-mark demonstrates that ALTE Member organisations aspire to the highest standards of quality and excellence in their exams.

The Q-mark indicates that the quality profile of an exam or suite of exams has been thoroughly audited by a professional ALTE auditor and the outcome of this audit was 'resolved'. The outcome remains valid for a period of five years or until there is a significant change in their validity argument for successfully audited exam(s). The Q-mark is only awarded to the audited exam(s), not to the organisation as a whole.

The Q-mark allows test users to be highly confident that an exam is founded on appropriate processes, reliable procedures and criteria, and consistent high standards.

Only the exams of ALTE Members and ALTE institutional affiliates applying for full membership, meeting certain criteria, can undergo an ALTE audit process and are therefore eligible to be awarded the Q-mark.

Full details of the exams which have been awarded the Q-mark, their registration numbers and dates for re-auditing are included in the ALTE Framework.

4. ALTE support and resources

4.1 Activities

ALTE's activities throughout the year support the primary aims of the association:

- **ALTE meetings**

  ALTE holds bi-annual meetings and conferences, with one conference day on a particular theme in language assessment open to the public. These are held usually in April and November, and are hosted by a Member organisation. Meetings include lectures, workshops on different aspects of language assessment, committee meetings, training opportunities as well as SIG group meetings. ALTE meetings are open to Members and Affiliates.

  From 2018 onwards, we make public a conference bibliography and the presentation slides of the plenary speakers and workshops, where possible.

- **SIGs (Special Interest Groups)**

  Within ALTE, Members can participate in special interest groups (SIGs), focusing on specific aspects within language assessment which are of interest to them. These include:

  - Common European Framework of Reference for Languages (CEFR)
  - QMS (ALTE Quality Management System)
  - Language Assessment for Migration and Integration (LAMI)
  - Young Learners (Teenagers and Children)
  - Teacher Training
  - Language for Specific Purposes
  - Special Requirements and Circumstances
  - Technology

  SIGs may also convene at an extra meeting, usually in February, which is also open to Members and Affiliates.

- **International conferences**

  Approximately every three years, ALTE organises an international conference, in cooperation with one of its Members. Experts in the field of language testing are invited as plenary speakers, on a general theme. This is an open conference, and papers may be presented on academic and applied aspects of language assessment. Recent conferences have been held in Bologna, Paris and Krakow, with papers given by world-leading experts on language testing. So far six very successful international conferences have been held, with the seventh happening in Madrid in April 2020. The proceedings of these conferences are available here.

- **Training**

  One of ALTE's aims is to promote assessment literacy, both for its Members and the wider educational community. To this end, different courses are provided, ranging from *ab initio* training for those new to the field, to highly specialised training on more technical aspects of language assessment – see Section 4.3, 'Services'. ALTE can also facilitate training for local Members where required, and training for those involved in the audit (see section 3, ALTE Quality Management System).

Further details of ALTE's past and future activities as well as conditions for participation can be found on the ALTE website: www.alte.org, or obtained from the ALTE Secretariat.

4.2 Materials

ALTE is offering guides and reference materials in a range of languages. They are all free to download from the Resources section on the ALTE website: www.alte.org/Materials.
Here is a selection:

- **Multilingual Glossary of Language Testing Terms**

  ALTE's multilingual glossary (1998) has a particularly significant role to play in encouraging the development of language testing in less widely taught languages by establishing terms which may be new alongside their well-known equivalents in the commonly used languages. The glossary contains entries in 10 languages: Catalan, Danish, Dutch, English, French, German, Irish, Italian, Portuguese and Spanish. This volume will be of use to many working in the context of European languages who are involved in testing and assessment.

- **Manual for Language Test Development and Examining**

  The *Manual for Language Test Development and Examining* (2011) was produced by ALTE on behalf of the Language Policy Unit of the Council of Europe. This manual is for use with the CEFR and it is available in Basque, Dutch, French and German.

- **Guidelines for the Development of Language for Specific Purposes Tests**

  The *Guidelines for the Development of Language for*

Specific Purposes Tests (2018) was produced by ALTE as a supplement to the 2011 *Manual for Language Test Development and Examining*. The production of the Guidelines was co-ordinated by the LSP SIG.

- **Relating Language Examinations to the CEFR: A Manual**

  ALTE contributed to the Council of Europe's Manual for *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)* (2009) and produced content analysis grids (2014) for Speaking and Writing (analysis and presentation). All of these are available on the Council of Europe's website.

- **ALTE Materials for the guidance of test item writers**

  This set of materials (2005) was designed to help in training anyone who is involved in any part of the process of developing, writing, administering and reporting the results of tests of a language learned as a foreign language.

- **Language tests for access, integration and citizenship: an outline for policy makers**

  This booklet (2016) was produced by the LAMI SIG on behalf of the Language Policy Unit of the Council of Europe. It is currently also available in Italian and Finnish.

- **ALTE Can Do Project**

  The ALTE 'Can Do' project developed and validated a set of performance-related scales, describing what learners can actually do in a language (2002). The project contributed greatly to the development of the CEFR and is acknowledged in Appendix D of the 2001 CEFR document.

  The Can Do Statements are available in the following languages:

  Català Dansk Nederlands English Suomi Français Deutsch Italiano Norsk Português Español Svenska

- **Content Analysis Checklists**

  Development and descriptive checklists for tasks and examinations (2001):
  General Checklist
  Reading Checklist
  Writing Checklist
  Listening Checklist
  Speaking Checklist
  Structural Competence Checklist

- **Checklists for Single Tasks**

  These content analysis checklists are for use with one task (2001):
  Single Reading Task Checklist
  Single Writing Task Checklist
  Single Listening Task Checklist
  Single Speaking Task Checklist
  Single Structural Competence Task Checklist

- **Quality Assurance Checklists**

  These checklists (2001) are matched to each stage of the exam production cycle:
  Quality Assurance Checklist 1 - Test Construction
  Quality Assurance Checklist 2 - Administration and Logistics
  Quality Assurance Checklist 3 - Marking, Grading & Results
  Quality Assurance Checklist 4 - Test Analysis and Post-examination Review

- **Bibliography of ALTE Members' work**

  ALTE has collated and keeps updating references of papers, journal articles, books, etc., written by representatives of ALTE Member organisations.

4.3 Services

4.3.1 ALTE courses

ALTE offers regular training opportunities across Europe for assessment professionals, teachers, and all those with an interest in the use and design of language exams. Besides the yearly ALTE Summer Courses, the Association also usually runs pre- and post-conference training sessions which are open to everyone. ALTE has also experience designing and delivering bespoke courses at the request of other organisations, both Members and external organisations. The courses are organised in three tiers, according to the increasing level of expertise required to participate. Listed here are courses that have been run in the past. This is not an exhaustive list, and other topics for courses can be considered:

- **Tier 1 – Introduction to language testing**

  These courses are intended for teachers and language professionals with limited knowledge about language testing theory and practice. Our most popular courses in this tier include:

  - Foundation Course in Language Testing: Getting Started
  - Introductory Course in Language Testing
  - Item Writing Training Course

- **Tier 2 – Language testing for professionals**

  Courses in this tier are aimed at language professionals who already have a good understanding of the principles and practice of language assessment, and would like to deepen their knowledge about certain aspects of language testing that are particularly relevant to their work. Examples of courses in this tier include:

  - Assessing Writing/Reading/Listening/Speaking Skills
  - Assessing Language for Academic Purposes
  - Validating Examinations with Fewer Candidates
  - Statistics for Language Assessment
  - Excel for Assessment Professionals
  - Introduction to Facets Analysis
  - Technology in Language Test Production and Validation
  - Managing Examinations for Language Testing Institutions

- **Tier 3 – Language testing for experts**

  Tier 3 includes specialised courses on technical issues and aspects of language testing. They are usually highly technical, and require a solid understanding and experience of language testing theory and practice. Some of our courses in this tier include:

  - Advanced Facets Analysis
  - Structural Equation Modelling for Language Testing
  - Differential Item Functioning

4.3.2 ALTE Validation Unit

The ALTE Validation Unit offers a range of services to exam providers and to small teams working in language assessment. This includes statistical analysis of exams and items, analysis of rater performance, reports on candidate characteristics, research into exam and language use at regional and national level, preparation for audits, editing and review of reports, and preparation of articles or presentations on assessment. Recent projects undertaken by the ALTE Validation Unit have included:

- Rasch and item analysis reports on the performance of items in pre-testing and live testing;
- analysis of rater performance in speaking and writing assessment;
- review of articles on training and monitoring of raters pre-submission to academic journals;
- research on the impact of introducing external assessment into national secondary education;
- pre-submission review of audit documentation for national exam providers;
- EU-funded research and a report on the use of the host country language in the workplace by migrant workers.

Requests for support from the ALTE Validation Unit can be placed with the ALTE Secretariat. Support is available for teams and individuals from both Member and external organisations.

Bibliography

American Educational Research Association, American Psychological Association and National Council on Measurement in Education (1985) *Standards for Educational and Psychological Testing*, Washington DC: American Educational Research Association Publishing.

American Educational Research Association, American Psychological Association and National Council on Measurement in Education (1999) *Standards for Educational and Psychological Testing*, Washington DC: American Educational Research Association Publishing.

American Educational Research Association, American Psychological Association and National Council on Measurement in Education (2014) *Standards for Educational and Psychological Testing*, Washington DC: American Educational Research Association Publishing.

Alderson, J C and Wall, D (1993). Does washback exist?, *Applied Linguistics* 14 (2), 115–129.

Bachman, L F (1991) *Fundamental Considerations in Language Testing*, Oxford: Oxford University Press.

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

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

Canale, M (1983) From communicative competence to communicative language pedagogy, in Richard, J C and Schmidt, R W (Eds) *Language and Communication*, London: Longman, 2–14.

Canale, M and Swain, M (1980) Theoretical bases of communicative approaches to second language teaching and testing, *Applied Linguistics* 1, 1–47.

Cheng, L (2005) *Changing Language Testing Through Language Teaching: A Washback Study*, Studies in Language Testing volume 21, Cambridge: UCLES/Cambridge University Press.

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

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

Fullan, M (1993) *Change Forces: Probing the Depths of Educational Reform*, London: Falmer Press.

Fullan, M (1999) *Change Forces: The Sequel*, Philadelphia: Falmer Press/Taylor & Francis, Inc.

Green, A (2003) *Test Impact and EAP: A Comparative Study in Backwash Between IELTS Preparation and University Pre-sessional Courses*, unpublished doctoral dissertation, University of Surrey.

Green, A (2007) *IELTS Washback in Context: Preparation for Academic Writing in Higher Education,* Studies in Language Testing volume 25, Cambridge: UCLES/Cambridge University Press.

International Language Testing Association (2000) *ILTA Code of Ethics*, available online: www.iltaonline.com

Joint Committee on Testing Practices (2004) *Code of Fair Testing Practices in Education*, available online: www.apa.org/science/programs/testing/fair-testing.pdf

Jones, N and Saville, N (2016) *Learning Oriented Assessment: A Systemic Approach*, Studies in Language Testing volume 45, Cambridge: UCLES/Cambridge University Press.

Messick, S (1980) Test validity and ethics of assessment, *American Psychologist* 35 (11), 1,012–1,027.

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

Milanovic, M and Saville, N (1996) *Considering the Impact of Cambridge EFL Examinations*, Cambridge: Cambridge ESOL internal report.

Saville, N (2009) *Developing a model for investigating the impact of language assessment within educational contexts by a public examination provider*, unpublished doctoral dissertation, University of Bedfordshire.

Toulmin, S E (2003) *The Uses of Argument*, Cambridge: Cambridge University Press.

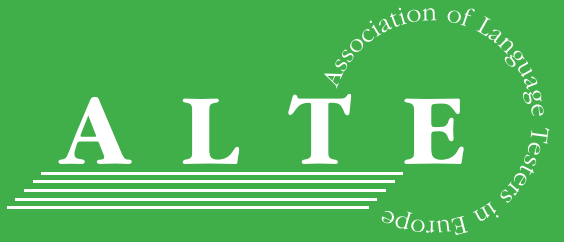
Van Avermaet, P, Kuijper, H and Saville, N (2004) A Code of Practice and Quality Management System for international language examinations, *Language Assessment Quarterly* 1 (2–3), 137–150.

Wall, D (2005) *The Impact of High-Stakes Examinations on Classroom Teaching: A Case Study Using Insights from Testing and Innovation Theory*, Studies in Language Testing volume 22, Cambridge: UCLES/Cambridge University Press.

Weir, C J (2005) *Language Testing and Validation: An Evidence-based Approach*, Basingstoke: Palgrave Macmillan.

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

Widdowson, H G (1983) *Learning Purpose and Language Use*, Oxford: Oxford University Press.


Association of Language Testers in Europe
ALTE