

The Association of Educational Assessment – Europe

# European Framework of Standards for Educational Assessment 1.0

## Contents

Foreword .....	3
1 The Framework of Standards for Educational Assessment .....	5
1.1 Definition .....	5
1.2 Purpose .....	5
1.3 Guiding Principles .....	5
1.4 Core Elements of Framework .....	9
1.4.1 Defining the Goal (construct, group, function) .....	9
1.4.2 Identifying the Nature of evidence and of tasks .....	10
1.4.3 Gathering evidence (admin and logistics) .....	10
1.4.4 Capturing outcomes (scoring, rating) .....	11
1.4.5 Decision making (aggregating, norms, grades, cut off scores) .....	11
1.4.6 Interpreting and reporting results .....	12
1.4.7 Evaluation and next iteration .....	12
1.5 Instrument .....	14
1.5.1 Defining the goal (construct, target group, function) .....	14
1.5.2 Identifying the Nature of evidence and tasks .....	15
1.5.3 Gathering evidence (admin and logistics) .....	16
1.5.4 Capturing outcomes (scoring, rating) .....	17
1.5.5 Decision Making (aggregating , norms, grades, cut off scores) .....	18
1.5.6 Interpreting and reporting results .....	19
1.5.7 Evaluation and next iteration .....	20
Annex A Context of Framework .....	21
1 Goals of AEA Europe .....	21
2 Definition of Framework of Standards .....	21
3 Purpose of Framework of Standards .....	22
4 Usage of Framework .....	23
5 European Dimension .....	24
5.1 Cooperation .....	24
5.2 Monitoring progress .....	25
5.3 Europe 2020 .....	26
Annex B Example of an review form .....	28
Annex C References .....	29

## Foreword

Are educational assessments and tests across Europe of the same quality? Are assessment and test developers in different European countries using the same quality principles? Can stakeholders find out in a transparent way how developers secure the quality of their assessments and is that quality comparable across Europe?

The Association for Educational Assessment Europe believes that currently the answers to these questions are not unequivocally positive.

The Association has the opinion that it is important that test developers provide information about the purpose of their assessments and about their claims of what the eventual scores will mean. So it asked itself the questions: Is it important therefore that there is some form of comparability across Europe in the quality notions we use and would such European comparability contain different aspects or approaches compared to existing concepts?

The European Framework of Standards for Educational Assessment presented here is the Association's answer to these questions.

The Framework is the result of a process that started a few years ago when AEA Europe members discussed these issues first during their annual conference and then in a more organized fashion. A Standards Committee was formed to come up with suggestions about if and how AEA Europe could play a role. In the 2009 conference this committee introduced the idea of producing a framework of European standards. After that a position paper was prepared that took into account input from AEA Europe members, examples from specific countries and documentation about standards and other quality assurances systems that are used in and outside Europe. This position paper was discussed during the annual conference of AEA Europe in Oslo, November 2010. Based on the feedback from that session, plus the other feedback received, the first draft of the Framework has been developed and the Association's Council adopted this as version 1.0 of The European Framework of Standards for Educational Assessment.

A key position that the Association has taken is to aim to produce a European Framework of Standards, instead of a European Set of Standards. The Framework is meant to be an instrument that test providers, score users and (educational) authorities can use to compare, contrast and evaluate practices in the development, administration, scoring and reporting of a wide variety of tests, assessments and assessment programs and that allows for the reviewing and presenting of evidence that the methods used meet certain quality criteria. It is first and foremost a tool that will facilitate transparency for providers, users and (educational) authorities. It can also be used for benchmarking an existing national system of standards for the development of assessment, for peer review activities and as a reference framework for the development of an audit procedure or an evaluation

framework to protect the rights of users and students. The presentation of this Framework is another step in realizing the long-term objective of the Council: to position AEA Europe as an expert organization that can facilitate and set up peer review and audit procedures to foster the quality of educational assessment across Europe.

Other benefits that might surface, once the Framework is better known and has been adopted across Europe, could be enhanced transferability of assessment results from one country to another in Europe. In addition such a framework could provide support for stakeholders in monitoring and evaluating the quality of assessments and assessment processes in their own setting, by identifying strengths and weaknesses, and by using feedback to inform improvements in their own practice.

The Framework is more comprehensive than a 'tool box'. It is not intended that it is technically detailed or strictly defines performance standards which must be followed. The Framework is communicative in nature and will establish itself through the accumulation of suitable methods for achieving standard requirements and best practises in documenting these methods. It will build up its practical applications, as people use it to review their own assessment systems, as it is used for peer review and as it increasingly forms the reference framework for audits.

We would like to thank Gerben van Lent, Andrew Watts and Saskia Wools for producing the first version of the Framework of Educational Standards for Europe and the other members of the Committee for their contributions and input: Kiril Bankov; Eduardo Cascallar; Jannette Elwood; Bas Hemker; Carolyn Hutchinson; Frans Kleintjes; Alastair Pollitt, and Gordon Stobart.

# 1 The Framework of Standards for Educational Assessment

This Framework document consists of the following parts:

1. Definition
2. Purpose
3. Guiding principles
4. Core elements
5. Instrument

## 1.1 Definition

The Framework is a practical instrument that test providers, score users and (educational) authorities can use to compare, contrast and evaluate practices in the development, administration, scoring and reporting of a wide variety of tests, assessments and assessment programs and to review and present evidence that the methods used meet certain quality criteria.

## 1.2 Purpose

The European Framework of Standards for Educational Assessment can be used by different assessment professionals for different purposes. Assessment developers might use the Framework to guide their activities, as it specifies the steps to undertake while developing a test. The Framework can also be used to evaluate existing assessment practices, as it suggests methods to be used and the evidence to be presented which is necessary for evaluation. It can be used for self-evaluation by test developers, it can guide peer-evaluation, or once fully developed it could be used for external evaluations and audits. Finally, the Framework can be used to compare educational assessment practice within Europe by providing a common reference and vocabulary to discuss assessment practice. This could extend to comparing the quality standards that are used within Europe. In all, the Framework aims to provide a European reference to facilitate greater transparency and comparability in the quality of assessment practices throughout Europe.

## 1.3 Guiding Principles

The Framework is based on five principles. It:

- *focuses on educational assessment*
- *is fit for a European environment*
- *emphasizes ethics / fairness and the rights of the individual*
- *addresses essential quality aspects: Validity, Practicality, Impact on stakeholders*
- *supports learning, decision making, test development, programme review.*

### **Focuses on educational assessment**

The Framework addresses those assessments which are used to record and support learning. The assessments to which the Framework applies cover most formal testing situations in assessment of learning as well as assessment for learning, including new forms of assessment. E.g.

- Standardised tests
- Examinations in various forms for a wide range of purposes
- School-based assessments (summative)
- Vocational / performance assessments and competency tests
- Assessment of systems / overall evaluation of total programmes or learning outcomes of a curriculum
- Innovative assessments, including in methods of delivery.

### **Fit for a European environment**

This is a key element of the Framework and its usage. It is AEA Europe's contribution of a European perspective to the world-wide interest in establishing standards as a way of disseminating quality in assessment. Europe has a wide range of approaches to assessment, and a variety of traditions of practice and research in the field of assessment. The Framework will have an integrating function and at the same time will support different cultural and educational contexts, as well as local definitions of 'fitness-for-purpose'.

### **Emphasizes ethics / fairness and the rights of the individual**

It is important to realize the students'/test takers' place in the assessment process. Assessment is an essentially human activity which aims to benefit learners and to enhance their life chances.

Therefore they can be viewed in the end as the main beneficiaries of what the Framework is intended to achieve. This focus on what is *good for the individual* aligns with the United Nations Convention on the Rights of the Child which directs that in all decisions made about children, their best interests must be "a primary consideration" and that children who can express a view have a right to have that view given "due weight" in all matters that affecting them. In addition, ethical considerations are now given more prominence in assessment processes than they have had in the past. The rights of test takers cannot be overlooked and best practice dictates that the assessment 'belongs'<sup>1</sup> to them and not to those who devise and administer it.

---

<sup>1</sup> 'This is meant in the sense of their being beneficiaries and not in a legal/formal sense (e.g. schools or authorities might want to claim ownership of an item bank and that would be problematic both for a testing organization and for the quality of the pool) or in a technical sense (e.g if test takers request full-disclosure of the items, then common-items could not be used for equating, or pre-test items, and a testing

### **Addresses essential quality aspects: Validity, Practicality, Impact on stakeholders**

These quality aspects of test and examination development are cornerstones of any professional assessment. The aim is to produce results which are meaningful and useful, and to show transparently how assessments translate the evidence that is gathered into results that are defensible. The statements about standards in the framework document will gain their credibility from being seen to be based on these fundamental assessment concerns.

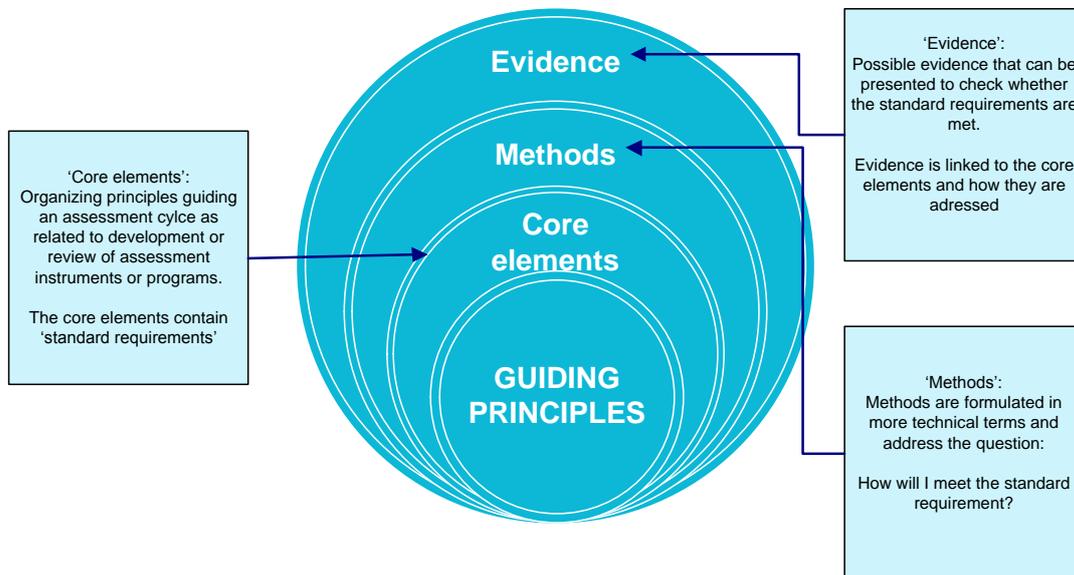
### **Supports learning, decision making, test development, programme review,**

AEA-Europe's fundamental aim is to support the education of learners. Assessments which are well-devised can enhance learning, whereas those that are poorly designed or unthinkingly applied can have a negative impact on students' learning progress. Educational assessment bases its rationale on the intended learning which underlies a particular educational process, and the Framework's aim is to enhance the experience of test takers so that assessments provide information that will invigorate them. The kinds of assessment envisaged include those that allow for feedback on the performance of educational systems to be used by decision makers to evaluate programmes and to allocate resources.

These guiding principles are at the heart of the Framework. From that starting point, seven core elements are distinguished which follow the assessment development cycle. What is expected for each of the core elements is described in the Framework in three levels. Each core element has standard requirements that need to be met. Then, for each core element, methods of implementation are described that are related to it. These methods describe *how* a standard requirement can be met. The next level is entitled 'evidence'. Here, the evidence is mentioned that can be presented as a backing for the other levels. Figure 1 presents this structure graphically.

---

organizations will have major difficulties in preserving the quality of the assessment and conducting equating of different test-forms. This will impact fairness.)



**Figure 1: structure of a European Framework of Standards for Educational Assessment**

## 1.4 Core Elements of Framework

The core elements which have been chosen for the Framework are derived from the development cycle of an assessment. Figure 2 below illustrates the development cycle that was followed.

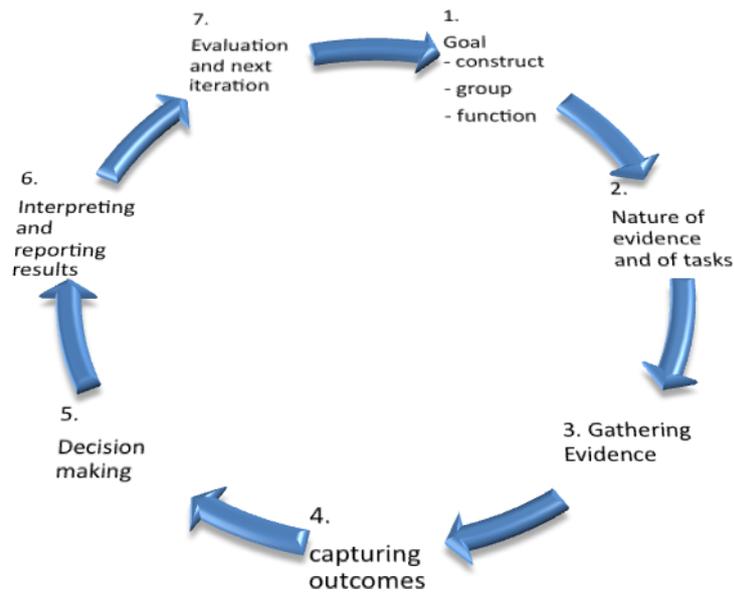


Figure 2: Development cycle and core elements.

### 1.4.1 Defining the Goal (construct, group, function)

A description of the goal of an assessment or assessment program contains:

- what a test, assessment or set of tasks is going to measure;
- what inferences are to be drawn from the results;
- who are the intended users of the results, and
- who are the intended test takers.

The goal is defined so that the content and constructs which the assessments are intended to measure, e.g. knowledge, skills, abilities, aptitudes or other characteristics, are linked to specific observable behaviours or performances which in turn are linked to tasks or situations that should elicit those behaviours. The chosen constructs are tied to explicit or implicit objectives of instruction, or are otherwise valued by the intended users.

The description of the goal indicates the possible score users and score use(s). This will allow for the rational development of construct-based scoring criteria and rubrics, and will support intended users in making quality decisions.

The candidate group for whom the assessment is intended is specified, detailing such aspects as age group, occupation, educational or occupational level.

The description states the strengths and limitations of the assessment, including the level of precision of the scores and the limitations of its use.

### **1.4.2 Identifying the Nature of evidence and of tasks**

The nature of the evidence to be collected and of the assessment tasks to be set flows directly from the definition of the goal of the assessment or assessment program. The combined tasks first and foremost should elicit the intended behaviours from the candidates in a fair, efficient, effective and engaging way. The assessment should be designed so that it has sufficiently representative coverage of the content and processes of the construct domain to allow reasonable claims regarding the extent to which the skills, knowledge and other attributes can be reproduced in settings other than the setting of the assessment. What is measured in the assessment should include important facets of the constructs that have been identified in the goal of the assessment (thus avoiding construct-under-representation).

Task formats should be designed that are fit-for-purpose and that do not contain elements that would discriminate amongst test takers in ways that are not the intent of the task. In other words 'construct-irrelevant variance' should be minimised.

Questions, items and tasks need to be combined with others to ensure that the assessment is stable, consistent and precise i.e. it provides repeatable outcomes for test takers with comparable characteristics (knowledge, skill, understanding) at different times and/or places. While discrimination is a fundamental purpose of the assessment (distinguishing those who are competent from those who are not) alongside discrimination can appear biases – unwanted effects in assessment which would give rise to unfairness. The development process should include sufficient internal and/or external reviews of the tasks and related materials.

Enough evidence must be collected to enable valid inferences to be made about test taker ability across the domain in question, taking into account the time available and the practicalities of conducting the assessments.

### **1.4.3 Gathering evidence (admin and logistics)**

In the gathering of evidence there are practical issues to be considered. The processes used for collecting evidence are important: for example, there must be no security leaks if the assessment is to be taken on sight. There must also be a level playing field so that test takers have the same chance to demonstrate what they know and can do. Any differences in their performances should be related to the construct which is the focus of the assessment.

On-line and internet assessment procedures to obtain the test takers' written outcomes should be preceded by proper identification procedures to guarantee each test taker's identity. In addition to collecting the evidence from the assessments, it is often relevant to collect *background variables* that relate to the outcomes. The distribution of background variables in the population(s) used for reference will be needed to evaluate the representativeness of outcomes.

#### **1.4.4 Capturing outcomes (scoring, rating)**

The purpose of this section is to ensure that valid decisions are made on the basis of the outcomes gathered. Proper scoring / marking must be separated from grading and come before it: the quality of the measurement instruments must be evaluated and assured, before measuring.

Those responsible for carrying out the assessments have a crucial role to play in ensuring the validity of assessments. If the outcomes depend on markers / scorers, such as is the case with open items in externally developed assessments, outcomes should be stored to facilitate *rater-agreement* studies. The selection, training, standardisation and supervision of the marking teams should ensure that a common standard has been set and that markers / scorers are applying that standard for the whole of the marking period.

In the case of a series of assessments, measures should be taken to store the outcomes on the various assessments in such a way as to allow the aggregation of results and to *validate overall decisions* based upon the combination of assessment outcomes.

Outcomes should be collected at *the lowest possible* level, e.g. if feasible, outcomes for individual candidate's answers should be collected on each item or task in the assessment, to facilitate proper evaluation.

The Framework of standards for this section applies to all types of assessment and delivery, including marking of *paper-based* and *computer-based* assessments.

#### **1.4.5 Decision making (aggregating, norms, grades, cut off scores)**

To reach a test taker's final score it might be necessary to combine the results of (sub)tests. To reach this score, different elements of information or evidence are combined into one decision on the assessment, for example when sum scores or weighted sum scores are used. It should be made explicit why a specific type of combination is selected, and what is the rationale for such a combination.

Subsequently, it will be necessary to give meaning to the (combined) scores or ratings by means of norms or performance standards. Several types of norms can be used: norm group-referenced, domain-referenced or criterion-referenced. Sometimes it is necessary to select more than one type of norm. In any case it should be made clear what type of decision is made, and how the chosen norms are relevant to that decision.

To achieve sound and reliable decisions it is necessary that appropriate procedures are used. E.g. to achieve performance standards, empirical evidence might be needed and thus relevant stakeholders should be represented within the assessment process to maintain performance standards. When candidates are compared with a norm group it is necessary to establish representative groups of a certain size. And where cut-off scores are chosen, it should be made clear why a certain cut-off score leads to the right decisions. For high-stakes standardized assessments with multiple test-forms it is necessary that the score comparability across test-forms is achieved, and that the variation in test form difficulty is adjusted. This comparability ensures that the test-takers who take different test forms are treated fairly.

#### **1.4.6 Interpreting and reporting results**

Procedures for interpreting assessment results should yield accurate and informative representations of test takers' performances in relation to the goals of the assessment. Proper interpretation of results should be guided by written policy or guidelines which incorporate explanations of results for students and all other relevant stakeholders, including some information about the uncertainty in results or decisions. This will allow suitable judgments to be made and appropriate decisions about who has access to the results (data protection policies).

Assessment reports should be clear, accurate and of practical value to test takers and, where applicable, to their parents/guardians and all other audiences for whom they are intended. Reporting systems should be at the appropriate level of detail, format, content, timing and confidentiality. Reports should also address the limitations and likely misinterpretations of the reporting scale. They should be guided by a documented policy which has been developed with input from developers and users of assessments.

When reporting results it is important to consider that personally identifiable scores are only published when candidates (or parents or guardians) have provided informed consent, except in circumstances in which consent is clearly implied, or when assessment without consent has been mandated by law or government regulation.

#### **1.4.7 Evaluation and next iteration**

The last step in the assessment process is to conduct an evaluation study and to use the results of this study to plan the next iteration of assessment. This next iteration might consist of the development of a new assessment or of the adapting and improving of an existing assessment. The evaluation study should at least focus on three elements: evaluation of technical aspects of the assessment, evaluation of the usefulness of the assessment, and evaluation of the impact of the assessment.

The evaluation of the technical aspects concerns analysis to determine the psychometric soundness of the assessment. The evaluation study of the usefulness should address issues like whether the process of administering the assessment was practical and efficient, and whether the reports were useful and informative. The **impact** of an assessment should be evaluated because all assessment practices will have an impact on learning and on the outcomes of education, shaping the behaviour of learners, teaching professionals and members of the wider community, including learner's families. The higher the stakes, the greater the impact will be on both learners and teachers.

## 1.5 Instrument

The following tables indicate how the three levels of the Framework work together. The headings in the table correspond with the core elements of the Framework as stated earlier. The *Standard Requirements* in the first column are directive. An assessment or assessment programme should address and meet all the Standard Requirements. The elements listed in the *Methods* and *Examples of Evidence* columns are illustrative and generic rather than prescriptive and specific. However, against each Standard Requirement the methods used need to be described, and for each method evidence has to be provided that is observable and verifiable.

### 1.5.1 Defining the goal (construct, target group, function)

Defining the Goal (construct, target group, function)		
What (Standard requirements)	Methods	Examples of evidence
<p>The goal of an assessment or assessment program should contain:</p> <ul style="list-style-type: none"> <li>• what an assessment, or set of assessments, is going to measure in terms of test takers' knowledge / skills / attributes</li> <li>• who are the intended test takers</li> <li>• what inferences can be drawn from the results</li> <li>• who are the intended users of the results.</li> </ul> <p>An explanation of the decisions taken about these elements should be included in the specification for the assessment.</p>	<ul style="list-style-type: none"> <li>• Identify the need for the assessment (for the individual, society, the economy) and refer to outcomes of curriculum or job analyses</li> <li>• Describe the reasoning behind the assessment, and define the construct(s) that the assessment measures</li> <li>• Justify the assessment content in terms of the construct to be measured</li> <li>• Consult stakeholders during the development process</li> <li>• Develop a specification which includes information about the content and skills or attributes to be measured; critical content to be included and content to be excluded; tasks to be performed; intended student population(s); interpretation of assessment results, and uses of those results</li> <li>• Ensure the assessment rationale reflects either an existing theory or is based on newly developed ideas that are described in sufficient detail by reference to the relevant literature</li> <li>• Conduct research about the construct validity and criterion validity of the assessment and make clear how the outcomes will support claims about what is being measured and will provide sufficient evidence to support the type of decisions for which the test is intended.</li> </ul>	<ul style="list-style-type: none"> <li>• Detailed specification document containing the key attributes of the assessment – what it measures, what inferences will be drawn, who are the intended candidates as well as users</li> <li>• Documentation that describes the process used to develop the specification</li> <li>• A rationale for the specification with references to other similar assessments and to relevant assessment literature</li> <li>• Research reports.</li> </ul>

## 1.5.2 Identifying the Nature of evidence and tasks

Identifying the Nature of evidence and tasks		
What (Standard requirements)	Methods	Examples of evidence
<ul style="list-style-type: none"> <li>The combined items, questions and tasks in an assessment should elicit the intended behaviours from the candidates in a fair, efficient, effective and engaging way</li> <li>The content covered in the assessment should include a sufficient range of the knowledge and skills associated with the important facets of the constructs that have been identified in the goal of the assessment</li> <li>Question and task formats should be designed that are fit for the required purpose and intended test takers</li> <li>Questions and tasks should be developed that meet generally accepted guidelines about what will be practical to administer, will produce results which can be validly and reliably used, and will have a positive impact on the learning of the candidates.</li> </ul>	<ul style="list-style-type: none"> <li>Develop tasks that are appropriate for the intended student group</li> <li>Develop assessment tasks that follow the given specifications</li> <li>Develop scoring rubrics and procedures, and training materials for scorers and/or automated scoring or score interpretation rules</li> <li>Identify how the assessment tasks are linked to the intended interpretations of assessment results</li> <li>Develop rules for assembling items, questions and tasks into an overall assessment</li> <li>Ensure the assessment represents the defined construct and excludes sources of variance unrelated to the construct</li> <li>Develop review procedures which require consideration of technical aspects of the assessment, its fairness to all test takers and of its language demand, appropriateness and fairness</li> <li>Provide evidence that the technical quality, including reliability and validity, of the assessment meets its intended purposes</li> <li>Provide guidance on the levels of skills, knowledge, and training necessary for appropriate development, review, and selection, of assessments</li> <li>Develop directions for students and ancillary materials to assist in briefing them.</li> </ul>	<ul style="list-style-type: none"> <li>Assessment and item/task construction specifications</li> <li>Descriptions of the assessment to help candidates prepare</li> <li>Documented description of the review process</li> <li>Technical reports on validity and reliability of the assessment</li> <li>Training manuals for assessment and item/task developers</li> <li>Information for users and students such as descriptive booklets and preparation materials.</li> </ul>

### 1.5.3 Gathering evidence (admin and logistics)

Gathering evidence		
What (Standard requirements)	Methods	Examples of evidence
<ul style="list-style-type: none"> <li>The forms of the evidence to be gathered (e.g. written examination scripts, recordings of speech, exemplars of project work) should be clearly described in the assessment specification</li> <li>The practicalities of gathering evidence should be developed in collaboration with the stakeholders to ensure that gathering of assessment evidence does not place unreasonable demands on test takers or administrators and should ensure that test takers will have an equal chance of demonstrating what they know and can do</li> </ul>	<ul style="list-style-type: none"> <li>Consult stakeholders about practicalities eg. paper storage, security, administration and despatch</li> <li>Where relevant, ensure test / examination centres have the required technical equipment</li> <li>In a user manual:               <ul style="list-style-type: none"> <li>specify the nature of the evidence that will be collected, e.g. exam scripts etc.</li> <li>ensure that the focus of each kind of evidence is clear to all involved</li> <li>explain how the assessment should be carried out in enough detail to ensure common conditions for all students</li> <li>ensure that the required level of security will be maintained.</li> </ul> </li> </ul> <p>Provide training for test administrators</p> <ul style="list-style-type: none"> <li>Ensure that the evidence is gathered in such a way as to facilitate the next step – i.e. scoring, rating</li> <li>Archive sufficient evidence to back validity claims and answer enquiries</li> <li>Where relevant, make sure the candidate’s identity is confirmed by some independent means</li> <li>Collect background variables in advance of the assessment application.</li> </ul>	<ul style="list-style-type: none"> <li>A specification which includes a justification of the amount of evidence required on which inferences will be based</li> <li>Pilot test or other reports on which time allocations have been decided, and later reviews based on feedback from administrations</li> <li>Where relevant, filed notes of meetings with and briefings for exam centres and administrators</li> <li>A manual for examination centres</li> <li>Feedback about administrative arrangements and reports from examination / testing centres</li> <li>A record of the contents of the assessment archives.</li> </ul>

### 1.5.4 Capturing outcomes (scoring, rating)

Capturing outcomes		
What (Standard requirements)	Methods	Examples of evidence
<ul style="list-style-type: none"> <li>• Marking / scoring / rating should produce results on which interpretations and decisions can be validly based</li> <li>• All those involved in marking / judgmental exercises should work to a common understanding of the required standards</li> <li>• Assessment data should be carefully collected and stored so that accurate studies of the candidates' performances and evaluations of the assessment tool itself can be carried out.</li> </ul>	<ul style="list-style-type: none"> <li>• Document the rationale and processes for marking / scoring in a way that is accessible to users Where judgments are required based on rating scales, e.g. for complex or for practical tasks, provide support in examples of the tasks and associated performances, which are annotated to explain how judgements were reached</li> <li>•</li> <li>• Carry out marking / scoring / rating of candidates' answers before grading decisions are made</li> <li>• Manage marking teams to ensure that a common standard is maintained by providing training and a system of quality control</li> <li>• To back validity claims, carry out <i>rater-agreement</i> studies</li> <li>• Set up data bases to allow outcomes for individual candidate's answers to be collected on each item or question</li> <li>• Carry out validation studies to confirm overall decisions based upon the combination of assessment outcomes.</li> </ul>	<ul style="list-style-type: none"> <li>• Reports of methods used for capturing evidence</li> <li>• Selection criteria for examiners / markers / raters</li> <li>• Training and standardisation materials</li> <li>• Records of examiner / rater / scorer performance</li> <li>• Outcomes of inter-rater agreement studies</li> <li>• Studies of the outcomes of an assessment, in terms of its effectiveness in providing for valid interpretations and decisions to be made.</li> </ul>

### 1.5.5 Decision Making (aggregating ,norms, grades, cut off scores)

Decision making (aggregating ,norms, grades, cut off scores)		
What (Standard requirements)	Methods	Examples of evidence
<ul style="list-style-type: none"> <li>• Meaning should be given to scores by means of norms. It should be made clear what type of decision has been made, and whether the types of norms are relevant to that decision</li> <li>• Procedures to achieve the norms should be carried out according to carefully established processes and documented.</li> <li>• When results of (sub)assessments are combined, it should be made explicit why a particular method of combination is selected</li> </ul>	<ul style="list-style-type: none"> <li>• Choose a method to aggregate scores and relate that to theory about the construct or content that is being measured</li> <li>• Elaborate the types of decision that will be made and explain any norms that are relevant to that decision</li> <li>• Follow standard setting procedures carefully</li> </ul> <p>Where relevant, establish representative norm groups of sufficient size</p> <p>Study whether any cut score produces valid results</p> <p>Report on how far the quoted reliabilities of the assessment are adequate for the purpose for which they will be used</p> <p>Provide background information about the members of an expert group, if applicable.</p> <p>The measurement methodology to be used to analyse the data from the assessment should be described in the specification</p>	<ul style="list-style-type: none"> <li>• Report on a study of the relationships within a construct</li> <li>• Manual which includes considerations about the norms</li> <li>• A description of norm groups and background information about the composition of the groups</li> <li>• Validation studies which include investigations of the relation to an external criterion to justify any cut off score</li> <li>• A description of any standard setting procedure followed.</li> </ul>

## 1.5.6 Interpreting and reporting results

Interpreting and reporting results		
What (Standard requirements)	Methods	Examples of evidence
<ul style="list-style-type: none"> <li>• Procedures should yield accurate and informative representations of performances</li> <li>• Interpretations should be guided by documented policy or guidelines which incorporate explanations of results and include possible ways of reflecting on and using assessment results</li> <li>• Assessment reports should be clear, accurate and of practical value to stakeholders</li> <li>• Reporting systems should be at the appropriate level of detail, format, content, timing and confidentiality</li> <li>• Consent should be given by test takers when identifiable scores or answers are reported to third parties.</li> </ul>	<ul style="list-style-type: none"> <li>• Make sure that the interpretations are aligned with the goals of the assessment</li> <li>• Ensure that the full scope of the decision is reflected within the final report</li> <li>• Describe the interpretations, the uses and the limitations of the results in a way that is understandable for both users and students</li> <li>• Develop guidelines with the help of key stakeholders, including students</li> <li>• Write reports in clear and understandable language for the target group</li> <li>• Investigate and assure the practical value of the reports</li> <li>• Design clear graphical elements and describe them so that they are informative for the target group(s)</li> <li>• Make sure that it is stated and argued what the appropriate level of detail, format, content, timing and confidentiality of reports are</li> <li>• Explicitly ask for consent when identifiable scores are reported and when this is not implied by the context of the assessment.</li> </ul>	<ul style="list-style-type: none"> <li>• Report on how the interpretation is aligned with the goals of the assessment with a paragraph added to the manual</li> <li>• Report on available information on the participants' performance which was used for the interpretation of the results</li> <li>• Consultation with experts or stakeholders to review reports for clarity</li> <li>• Studies of whether users find the reports of practical value</li> <li>• Review by members of the target group(s) of the clarity of the reports</li> <li>• Studies of whether users find the graphical elements of reports clear.</li> </ul>

## 1.5.7 Evaluation and next iteration

Evaluation and next iteration		
What (Standard requirements)	Methods	Examples of evidence
<ul style="list-style-type: none"> <li>• Evaluation studies of assessments should be conducted and the results of these studies used to plan the next iteration</li> <li>• The technical aspects of assessments should be evaluated to judge the psychometric soundness of the assessment</li> <li>• The usefulness of the assessment to its stakeholders should be evaluated</li> <li>• The impact of the assessment should also be evaluated, particularly in so far as it affects the candidates' learning.</li> </ul>	<ul style="list-style-type: none"> <li>• Ensure that the results of the evaluation study are used to improve the next iteration</li> <li>• Decide how an assessment can be improved based on the results of the evaluation study</li> <li>• Decide how the development process that leads to a new assessment can be improved based on the results of the evaluation study</li> <li>• Use psychometric/statistical analyses to establish the technical quality of the assessment</li> <li>• Study the experiences of students, teachers and other involved stakeholders</li> <li>• Perform a study that addresses issues like the practicality or efficiency of the administration of the assessment</li> <li>• Evaluate whether the results were useful and informative</li> <li>• Gather views of all those with a stake in education about the impact of different assessment and reporting arrangements</li> <li>• Investigate how the assessment affects what happens in educational establishments and classrooms</li> <li>• Perform an impact study that addresses whether the assessment supports the curriculum, motivates students, and promotes deep learning.</li> </ul>	<ul style="list-style-type: none"> <li>• Research reports of evaluation studies of the assessment instruments</li> <li>• Technical reports of the assessment results</li> <li>• Transcripts of interviews on the experiences of students and teachers</li> <li>• Research reports on the impact of the assessments</li> </ul>

## **Annex A Context of Framework**

### **1 Goals of AEA Europe**

The general goal of AEA Europe is to act as a platform for discussion of developments in educational assessment in Europe by fostering co-operation and facilitating liaison between organisations and persons active in educational assessment across the whole of Europe. AEA Europe defines educational assessment in its broadest sense to include assessment in academic, professional and vocational contexts and it is equally concerned with both assessment processes and products.

The aims of AEA Europe include:

- To engage in a range of activities that will lead to the improvement of assessment processes and products and their appropriate use by organisations, institutions, agencies and other associations throughout Europe;
- To enhance awareness of assessment processes and products in relation to their impact on learning and understanding.

The role of the Association is to help co-ordinate efforts to improve assessment systems and practice throughout Europe, through contact between organisations involved in examinations and assessment research and individuals working professionally in this area. These contacts promote scientific knowledge and the application of that knowledge to practical problems in school systems.

The membership emphasizes the importance of assessment as being at the heart of the educational process. Good assessment practices are vital to the promotion of learning, both through classroom assessment and through valid examination systems. Most educational reform programmes and projects recognise the importance of assessment in the structures they create. For these reasons the Association has concluded that, in addition to its basic activities, it should promote the importance of assessment and seek the active support of educational policy makers, and their supporting funding organisations, in all parts of Europe.

In its recent strategic plan the Association formulated its mission as: “To be the association to which assessment professionals throughout Europe look to for information, contacts and career development.”

To deliver this mission, the Association identified introducing a Framework of European Standards for Educational Assessment as one important goal to increase its authority and standing in the professional community, so that people may wish to join and contribute.

### **2 Definition of Framework of Standards**

The Framework is an instrument that test providers, score users and (educational) authorities can use to compare, contrast and evaluate practices in the development, administration, scoring and

reporting of a wide variety of tests, assessments and assessment programs and that allows for the reviewing and presenting of evidence that the methods used meet certain quality criteria. The Framework sets out key principles of assessment and it describes the principles in operational terms. It is designed to apply to a wide range of assessments, including high-stakes traditional examinations, vocational assessments, large-scale assessment programmes, national systems and one-off tests.

A key position that the Association took was to aim to produce a European Framework for Standards, instead of a European Set of Standards.

In this document the word 'standards' refers to quality standards: when in a specific part of the text performance standards are meant, it will be mentioned explicitly.

### **3 Purpose of Framework of Standards**

The European Framework of Standards for Educational Assessment will establish a practical common European reference which will link various national and professional assessment standards systems and practices together and so facilitate greater transparency and comparability of these systems and practises. A network of independent but related and mutually understandable national and professional assessment standards systems will thereby be created. The Framework uses Core Elements of an assessment cycle to describe the development or review of assessment instruments or assessment programmes. As a common reference point, the Framework will facilitate comparison and transfer of standards and best practices between countries, systems and institutions and will therefore be relevant to a wide range of users at European as well as at national level.

This closer relationship between national and professional assessment standard systems and practices will have many benefits:

- The European Framework of European Standards for Educational Assessment will support individual users, as well as providers of education and training, by increasing the transparency of certificates and credentials awarded inside and outside national systems. This will help build the trust of authorities and employers who will assess the value of qualifications of applicants and so support education and labour market mobility in Europe.
- The Framework will have a trickledown effect and eventually might make it easier for learners to provide evidence and credentials, reflecting their level of competence, to educational institutes and or company recruiters in their own or other countries, because the evidence they provide is based on shared understanding of best practises.
- At a system level the Framework should be seen as supporting Quality Management systems for assessment policy-makers and developers. It will be a tool, which could be used to elicit evidence

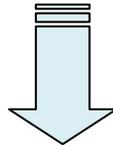
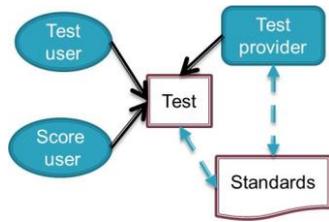
that could drive organisation change or develop arguments for allocation of resources and funding.

## **4 Usage of Framework**

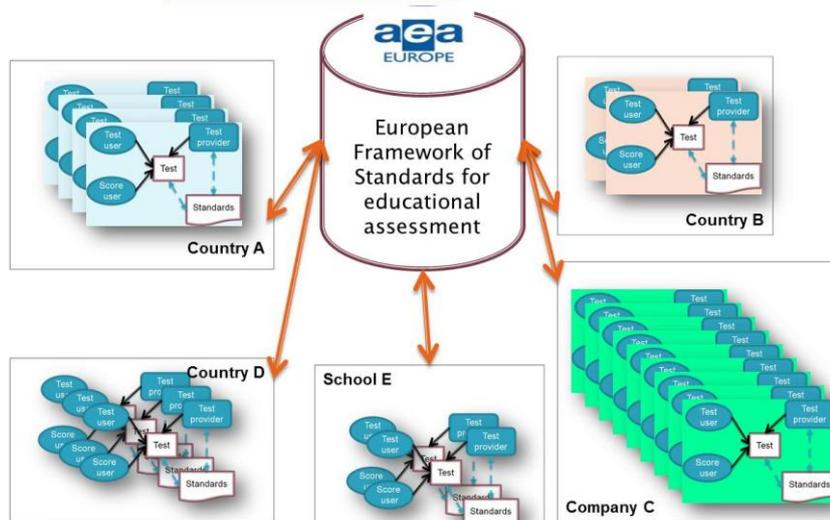
The Framework is first and foremost a tool that will allow for transparency for providers, assessment users and (educational) authorities. It could be used for benchmarking an existing national system of standards for the development of assessment and for peer review activities. As a reference framework, it could also be used for the development of an audit procedure or an evaluation framework to protect the rights of users and test takers. The Framework will enhance understanding of educational assessment and support stakeholders to monitor and evaluate the quality of assessments and assessment processes in their own setting, identifying strengths and weaknesses, and using feedback to inform improvements in their own practice. It should enable stakeholders to evaluate the fairness and transparency of decisions taken on the basis of assessments and to judge the way that results are reported and used. The Framework could serve to guide the development of new approaches to assessment which are fit-for-purpose and which meet the technical standards which underpin quality in assessment. In time AEA Europe hopes, as an expert organization for assessment in education, to facilitate and set up peer review and audit procedures to foster the quality of educational assessment across Europe.

The envisioned usage of the Framework in local settings is as follows. The first step would be to use it as a tool for self-evaluation of quality. Individuals and organizations could use the Framework to evaluate their own assessment practice and to identify strengths and weaknesses. The results of this self-evaluation will enhance assessment quality due to the insights gained during the evaluation. In the second phase, the Framework could serve as a tool for peer review. Both individuals and organizations could evaluate their assessments together or ask for a review from a colleague or from AEA Europe. The Framework would guide the feedback that is given on assessments and provide the opportunity to start discussions between practitioners about their local assessment practice and what they can improve. The Association could act as a facilitator of this process.

In future, when the framework is fully developed and used for self evaluation as well as peer reviews, it is the intention of the AEA Europe to set up a coordinating body. This coordinating body may, for example, certify certain assessments or assessment practices when they comply with the Framework. It might also be possible for users or stakeholders to address this coordinating body concerning complaints or doubts about a certain assessment or assessment procedure and to ask whether the assessment is sound.



## European Framework of Standards for Educational Assessment



## 5 European Dimension

The development of a European Framework for Standards for Educational Assessment fits well into the strategies that the EU Member States and the European Commission have implemented in the past decade and will continue to implement through the recently published Europe 2020 strategy. Below, an overview is provided of key developments and initiatives in the past decade and in the decade to come. This is followed by a brief rationale for why the Framework supports these developments and initiatives.

### 5.1 Cooperation

In the updated strategic framework for European cooperation in education and training (Communication from the Commission, COM(2008) 865, 16 December 2008) it was stated that European education and training systems need reform to better prepare people to find jobs and to

help businesses find the staff they need to succeed and innovate in the face of global competition. The quality and efficiency of education systems play a key role in supporting social inclusion, cultivating responsible citizenship and openness towards other cultures. For that reason, lifelong learning must become a reality across Europe so people can acquire key skills early and update them throughout their life.

While the responsibility for education and training lies with national governments, some challenges are common to all EU countries. Therefore EU Member States and the European Commission have in recent years strengthened their co-operation in education and training at the European level. *The focus is on exchanging good practice and developing common tools, in order to face common challenges.* In support of future reforms in the Member States this new framework will strengthen mutual learning between them.

## **5.2 Monitoring progress**

Another element that EU member States and the European Commission value is the gathering of data and the collection of evidence to monitor initiatives.

In 2006 the Council of Europe claimed: “Considerable progress has been achieved through cooperation so far - particularly in support of national reforms of lifelong learning, the modernisation of higher education *and the development of common European instruments* promoting quality, transparency and mobility.” The Council's key messages concerning the field of education and training to the Spring 2007 European Council emphasised that “educational policies and practices require a stronger evidence base” and that “there is a need to develop a culture of evaluation and research”. This was followed by the Council’s conclusions of 25 May 2007 on a coherent framework of indicators and benchmarks for monitoring progress towards the Lisbon objectives in education and training: “... *the development of new indicators shall fully respect the responsibility of Member States for the organisation of their education systems and should not impose undue administrative or financial burdens on the organisation and institutions concerned*”. Periodic monitoring of performance and progress through the use of indicators and benchmarks is an essential part of the Lisbon process, allowing strengths and weaknesses to be identified with a view to providing strategic guidance to the ‘Education and Training 2010’ work programme.

It was noted that considerable progress has been achieved through cooperation so far - particularly in support of national reforms of lifelong learning, the modernisation of higher education and the development of common European instruments promoting quality, transparency and mobility.

### **5.3 Europe 2020**

The European Commission presented the Europe 2020 Strategy to help Europe emerge from the financial crisis and prepare its economy for the next decade. In the executive summary the following is stated:

“Education and training are to play a key role for smart, sustainable and inclusive growth.”

The Europe 2020 Strategy, the follow-up to the former Lisbon-Strategy, will focus on a trio of priorities which are linked and will reinforce each other:

- **Smart growth:** developing an economy based on knowledge and innovation.
- **Sustainable growth:** promoting a more resource efficient, greener and more competitive economy.
- **Inclusive growth:** fostering a high-employment economy delivering social and territorial cohesion.

Education, training and lifelong learning play a key role to achieve these strategic priorities, in particular when it comes to smart and inclusive growth. In order to meet the set priorities and targets, the Commission proposes a Europe 2020 agenda consisting of a series of flagship initiatives. Implementing these initiatives is a shared priority, and action will be required at all levels: EU-level organisations, Member States, local and regional authorities, and civil society.

Two of these flagship initiatives will be particularly linked to education and training:

*Youth on the move:* the aim is to enhance the quality and international attractiveness of Europe's higher education system by promoting mobility of students and trainees. As a concrete action, the EU's mobility programmes should be enhanced and linked to national programmes, the modernisation agenda of higher education pursued and professional experience should be properly recognised.

*An agenda for new skills and jobs* has the objective of creating the right conditions to modernise labour markets and to allow people to acquire new skills in order to raise employment levels and to ensure the sustainability of our social models, while baby-boomers retire. The concrete actions to be taken include the implementation of the strategic framework for European cooperation in education and training, the acquisition and recognition of learning throughout general, vocational, higher and adult learning, and the implementation of the European Qualifications Framework.

It is interesting to note that common instruments that have been developed to date have focussed on outcomes of education, and achieved knowledge, skills and attitudes of learners e.g. the CEFL and the EQF, or on transparency, quality and mobility across Europe through the Bologna and Lisbon Agreements. While gathering data about the outcomes of learning and the credentialing of competencies have been recognized as major drivers for implementing the various strategies,

relatively little attention at a European level has been given to how (the quality of) tests, examinations, assessments and assessment programs used across Europe compare to one another. It is against this background that the Association for Educational Assessment Europe has initiated the development of the European Framework of Standards for Educational Assessment. Respecting the traditions within the EU and the division of responsibilities within national and EU cooperation, the Framework is an instrument to be used to compare and contrast practises in assessment.

## Annex B Example of an review form

Standard requirement	Methods	Evidence	Comment
<b>1. Defining the goal</b> The goal of an assessment or assessment program should contain four key elements:	<ul style="list-style-type: none"> <li>Identify the need for the assessment (for the individual, society, the economy) and refer to outcomes of curriculum or job analyses</li> <li>Describe the reasoning behind the assessment, and define the construct(s) that the assessment measures</li> <li>Justify the assessment content in terms of the construct to be measured</li> <li>Consult stakeholders during the development process</li> <li>Develop a specification which includes information about the content and skills or attributes to be measured; critical content to be included and content to be excluded; tasks to be performed; intended student population(s); interpretation of assessment results, and uses of those results</li> <li>Ensure the assessment rationale reflects either an existing theory or is based on newly developed ideas that are described in sufficient detail by reference to the relevant literature</li> <li>Conduct research about the construct validity and criterion validity of the assessment and make clear how the outcomes will support claims about what is being measured and will provide sufficient evidence to support the type of decisions for which the test is intended.</li> </ul>	<ul style="list-style-type: none"> <li>Detailed specification document containing the key attributes of the assessment – what it measures, what inferences will be drawn, who are the intended candidates as well as users</li> <li>Documentation that describes the process used to develop the specification</li> <li>A rationale for the specification with references to other similar assessments and to relevant assessment literature</li> <li>Research reports.</li> </ul>	
What an assessment, or set of assessments, is going to measure in terms of test takers' knowledge / skills / attributes			
Who are the intended test takers			
What inferences can be drawn from the results			
Who are the intended users of the results.			

## Annex C References

### Comparison with existing sets of Standards

WHAT?	ALTE Quality profiles	Dutch rating system	APA testing practices	International Guidelines for Test Use	ETS standards	Psychological tests: EPPA review model
<b>Goal: Construct, group, function</b>	1-5 Test construction	1 Theoretical basis 5A Construct validity	P4,5 Developing and selecting tests	2.1 Use of tests	Chaps 2 & 12 Suitability for use Ch. 6 Validity	Section 2 Classification: contents, purpose, design
<b>Tasks and evidence, nature of</b>		2 Test materials and manuals	P4 Test developers	2.2 Technically sound tests 2.3 Fairness	Ch. 7 Assessment development Ch 6 Validity	Section 7 Validity
<b>Gathering evidence (admin and logistics)</b>	6-10 Admin and logistics	2B Test manual	P6&7 Administering and scoring	2.4 Preparations 2.5 Administration	Chap 10 Assessment administration Ch 4 Fairness	Section 5 Supply, condition and costs
<b>Capturing outcomes (scoring, rating)</b>	11,12 Marking and grading	2B Manual 4 Reliability 5B Criterion validity	P6 Test developers	2.6 Scoring	Chap 8 Reliability	Section 3 Measurement and scoring Sect. 7 Reliability
<b>Decision making (aggregating, norms, grades, cut off scores)</b>	13,14 Test analysis	3 Norms	P8 Interpreting results	2.6 analysing results	Chap 9 Cut scores, scaling, equating	Section 7 Evaluation of Norms, reliability and validity
<b>Interpreting and reporting results</b>	15,16,17 Communication with stakeholders	2, 9-13 Interpreting scores	P8&9 Reporting and interpreting results P10 Informing test takers	2.7 Interpreting results 2.8 Reporting results	Chap 11 Reporting assessment results	Sections 4 & 7 Reports
<b>Evaluation and next iteration</b>				2.9 Reviewing appropriacy of test	Chap 1 Development procedures	Sections 6 & 8 Evaluation of test materials