EALTA
www.ealta.eu
European Association for Language Testing and Assessment

EALTA GUIDELINES FOR GOOD PRACTICE IN LANGUAGE TESTING AND ASSESSMENT

The EALTA Mission Statement is as follows:

> The purpose of EALTA is to promote the understanding of theoretical principles of language testing and assessment, and the improvement and sharing of testing and assessment practices throughout Europe.

The rationale for the Guidelines is as follows:

Europe is a multilingual continent, where the diversity of languages, cultures and traditions is highly valued. Such diversity is represented in education systems and assessment traditions and values.

Consequently, EALTA members strive to adhere to the shared overarching principles of transparency, accountability and quality in ways that are appropriate to their particular contexts and spheres of professional involvement in language testing and assessment.

Reflecting its policy of inclusiveness, EALTA wishes to serve the needs of a broad membership. EALTA's guidelines for good practice in testing and assessment are accordingly addressed to members involved in one or more of the following three activities: (a) the development of tests for certification recognised at a national or institutional level, (b) the pre- or in-service training in assessment of language teachers, and (c) classroom-based testing and assessment.

For all these activities, a number of general principles apply: respect for the students/examinees, responsibility, fairness, reliability, validity, and collaboration among the parties involved. These general principles are laid out in relevant existing codes of practice, which EALTA members are encouraged to consult in order to further the professionalism, quality and beneficial consequences of their work. Please refer to the following links:

http://www.iltaonline.com
http://www.alte.org

EALTA's own guidelines for good practice in language testing and assessment, designed to be useful and appropriate for the European context and common professional activities of EALTA members, are as follows:

A. Considerations for test development for certification recognised at an international, national or local level

EALTA members involved in test development clarify to themselves and appropriate stakeholders (teachers, students, parents/guardians, test score users, the general public) by providing answers to the questions listed under the headings below. Furthermore, test developers are encouraged to engage in dialogue with relevant decision makers in educational, employment and governmental institutions to ensure that decision makers are aware of both good and bad practice, in order to enhance the quality and beneficial consequences of assessment systems and practices.

1. TEST PURPOSE AND SPECIFICATION

1. How clearly is/are test purpose(s) specified?
2. To what extent are all stakeholders specifically identified?
3. Are there test specifications? Are they (or some version of them) available publicly?
4. To what extent are the specifications for the various audiences differentiated?
5. How clearly are the test takers described?
6. How clearly are the constructs intended to underlie the test/subtest(s) specified?
7. How clearly are test methods/tasks described and exemplified?
8. To what extent is the range of student performances described and exemplified?
9. How clearly are marking schemes/rating criteria described?
10. If test levels specified in CEFR terms, what evidence is provided to support the linking to the CEFR?

2. TEST DESIGN and ITEM WRITING

1. What measures are taken in test design to ensure that the test is accessible to all test takers in the target population?
2. What measures are taken to ensure that test content is representative of different groups in the test-taking population and TLU domain?
3. To what extent are relevant stakeholder groups, such as different groups of test-takers and their teachers, consulted when new test formats are introduced/developed?
4. How detailed are the guidelines for test design and item writing?
5. What are the procedures for review, revision and editing of items and tasks to ensure that they match the test specifications and comply with item writer guidelines?
6. What feedback do item writers receive on their work?
7. Does the development of a new rating scale involve different (relevant) stakeholders?
8. To what extent are the properties of new rating scales investigated by various quantitative and qualitative procedures?
9. What measures are taken to ensure that the test developers and item writers have relevant knowledge and/or experience of the targeted level and the context of the assessment?
10. What training is provided to the test developers and item writers?

11. If AI is used, can its role in the item writing and/or content generation process be described? Can it be described how its use is monitored?

3. QUALITY CONTROL and TEST ANALYSES

1. What quality control procedures are applied?
2. Are the tests piloted on a sufficiently large and representative sample of test takers?
3. How is the evidence from the pilot test analyses used and acted upon?
4. If there are different forms or (amended) versions of the test (e.g., year by year), how is their equivalence/comparability verified?
5. How are pass/fail grades and other cut scores denoting key decision points determined?
6. To what extent are markers trained for each test administration?
7. To what extent are benchmarked performances used in the training?
8. How is the accuracy of subjectively marked tests ensured?
9. To what extent is the (human and/or machine) marking monitored?
10. What statistical post-test analyses are used (e.g., for investigating item and test characteristics, bias, rating quality)?
11. What results are reported? How? To whom?
12. What processes are in place for test takers to make complaints or seek re-assessments?

4. TEST ADMINISTRATION

1. What administrative procedures are in place to reduce threats to test security and address reports of potential malpractice?
2. How are test administrators trained?
3. How is the test administration monitored?
4. What accommodations are provided to ensure that the test is accessible to test-takers with different needs?
5. Is there a post-administration report at regular intervals?

5. REVIEW

1. If the test is based on a curriculum or a domain (the real-world context in which the language being assessed is used), what procedures are in place to ensure that the test keeps pace with changes in the curriculum/domain?
2. If the test is based on a theory/model of language proficiency, what procedures are in place to ensure that the test keeps pace with developments in such theories/models?
3. To what extent is test-taker feedback considered in the review process?
4. To what extent are validation studies conducted?

6. WASHBACK (IMPACT ON TEACHING AND LEARNING)

1. What is the washback effect of the test? What studies have been conducted?

2. To what extent are there preparatory materials?
3. To what extent are there guidelines/material that support teachers in how to use the tests formatively/to enhance learning?
4. To what extent are the links between the curriculum and/or target language use domain and test content coverage made clear?
5. To what extent are teachers trained to prepare their students for the test/exam?
6. To what extent have learners had sufficient opportunity to acquire the knowledge and skills being tested?

7. TEST USE AND CONSEQUENCES

1. What measures are taken to ensure that test scores are interpreted and used in line with the intended purpose(s)?
2. To what extent is information collected about how the test is used by different stakeholders/policy makers?
3. What measures are taken if test misuse is detected (e.g., if requirements for access to education, labour, residence rights are set too high or a test is used for a purpose for which it was not intended)?
4. To what extent are the consequences of the use of the test(s) on different test-taker groups investigated?
5. What measures are taken if negative consequences for test takers or society are detected?

8. LINKAGE TO THE COMMON EUROPEAN FRAMEWORK (CEFR)

Many language assessment systems find it useful to link their assessments to the CEFR, particularly in Europe. If a particular linkage is claimed, a list of the steps needed in the linking process is found in section D of this document, which also includes links to useful external sources about this topic. Note, however, that not all assessments need to be linked with the CEFR.

B. Considerations for the pre-service or in-service training in assessment of language teachers

EALTA members involved in teacher education related to testing and assessment clarify to themselves and appropriate stakeholders (trainees, practising teachers, curriculum developers):

Components foregrounding theoretical concepts

1. To what extent does the training engage the trainees' knowledge about more general linguistics theories and schools of thought (e.g., the role of language in our societies, how languages are learned)?

2. How clearly are the principles of testing and assessment (e.g., validity, reliability, washback, fairness, justice) related to the trainees' context?
3. How relevant is the training to the local or national assessment context of the trainees?
4. To what extent are trainees aware of the range of the different types of assessment they may engage with?
5. To what extent are trainees aware of the relationship between language teaching, learning and assessment?

Components foregrounding practical work

6. How are trainees supported in linking assessment theory to practice in the training?
7. To what extent are the technical skills addressed by the training (e.g., developing specific rating scales, applying rating scales, or developing/reviewing items or tasks) relevant to the trainees' present or future needs as language teachers?
8. What attention is given to formative assessment and how to give meaningful feedback to learners?
9. What attention is given to the appropriate analysis and interpretation of assessment results from classroom assessments or standardised examinations?

Components foregrounding ethical considerations

10. To what extent are trainees given room to share their views on the appropriacy and accuracy of assessment procedures?
11. To what extent are trainees made aware of the social consequences of tests on different groups of test takers?
12. Depending on the local or national context, how are trainees made aware of the affordances and limitations of using generative language models and technology in language testing and assessment?

C. Considerations for classroom testing and assessment

EALTA members involved in classroom testing and assessment clarify to themselves and appropriate stakeholders (especially pupils/students and, where appropriate, parents/guardians):

1. ASSESSMENT PURPOSE(S) AND SPECIFICATION

1. What is the purpose of the assessment?
2. To what extent are decisions made on the basis of the test results?
3. How does the assessment purpose relate to the curriculum?
4. To what extent are a range of assessment methods (besides tests) to evaluate learners' ability employed?
5. Are there any test specifications? To what extent/in how much detail do you plan your assessments (and record this)?
6. How well is the curriculum covered?

7. How are the assessment purposes and specifications made known and discussed with students?

2. ASSESSMENT PROCEDURES

1. How are the assessment procedures decided?
2. How appropriate are the assessment procedures to the learners?
3. Do you use AI in assessment? If you use it, can you describe how it is used, justify its use, and describe how its use is monitored?
4. How do you collect information on students' learning?
5. How do you assess and store information on students' learning?
6. What efforts do you make to ensure that the assessment results are accurate and fair?
7. What efforts do you make to promote agreement in marking practices across teachers and schools?
8. How do you take into account students' views on the assessment procedures?
9. What accommodations are provided to ensure that the test is accessible to test-takers with different needs?

3. WASHBACK

1. What use do you make of the results? (e.g., decide on the grades based on learner achievement or to provide learners with formative feedback)
2. What action(s) will you take to improve learning?
3. What kind of feedback do you give to students?
4. What processes do you have in place for students or their guardians to make complaints or seek re-assessments?
5. Do you consider the consequences of the assessment procedures for classroom practices?
6. Do you consider the consequences of the results of the assessment for learners?

D. LINKAGE TO THE COMMON EUROPEAN FRAMEWORK (CEFR)

Linking to the CEFR is a complex endeavour, which may often take the form of a project to be developed over a number of years. Linking exams or tests to a standard such as the CEFR requires a scientific approach, and claims must be based on the results of research, preferably submitted for peer review.

Institutions/exam providers wishing to claim linkage to the CEFR are accountable for the provision of sufficient convincing evidence for such linkage. The following considerations may be useful to gather such evidence. More comprehensive guidance on how to align assessments (and language education more generally) with the CEFR can be found in the handbook designed by EALTA, ALTE, UKALTA and the British Council Aligning language education with the CEFR. The handbook can be accessed in the Resources section on the EALTA website.

EALTA members involved in the process of linking an exam to the CEFR will clarify to themselves and appropriate stakeholders and provide answers to the questions below:

1. Have issues of test quality and impact been fully addressed? (see EALTA Guidelines Section A). Is information on the quality of the test publicly available?

2. Are those involved in test development familiar with the CEFR? Have CEFR familiarisation procedures been followed? What evidence is there of the procedures followed and the corresponding results?

3. Have the test content and test specifications been analysed in relation to the CEFR descriptors? What evidence is there of the procedures followed and the corresponding results?

4. Have standardisation procedures been completed for written and spoken performances? Which? By whom? What evidence is there of the procedures followed and the corresponding results?

5. Are local performance samples that are benchmarked to the CEFR publicly available?

6. Have standardisation procedures been completed for test items targeting receptive skills? Which? By whom? What evidence is there of the procedures followed and the corresponding results?

7. Are local CEFR-linked items and tasks publicly available?

8. What standard setting procedures have been used to establish cut-off scores for the relevant CEFR level(s)? Is a standard setting report publicly available?

9. Has validity evidence been collected covering the process of linkage? Is it publicly available?

HISTORY OF THE EALTA CODE OF PRACTICE

The original Code of Practice was drafted by the ENLTA (European Network for Language Testing and Assessment) Project in 2003-2005 that laid the groundwork for the establishment of EALTA. The EALTA Executive Committee appointed a Working Group to review and revise the draft Code of Practice and to develop a code of practice or set of guidelines for good practice in testing and assessment which is appropriate to EALTA and its mission. In 2024, the EALTA Executive Committee set up a Working Group to revise the Code of Practice in light of the developments in the field over that past two decades.