International Language Testing Association

Guidelines for Practice

2024

Suggested citation: International Language Testing Association. (2024). ILTA Guidelines for Practice. Available at: https://www.iltaonline.com/page/ILTAGuidelinesforPractice

Contents

1 OVERVIEW OF THE ILTA GUIDELINES 3

1.1 WHO ARE THE GUIDELINES FOR? 3
1.2 WHAT IS COVERED IN THE GUIDELINES? 3
1.3 HOW TO USE THE GUIDELINES 4

2 SUMMARY OF KEY CONSIDERATIONS 5

2.1 KEY CONSIDERATIONS IN DESIGNING AND USING FORMAL LANGUAGE TESTS 5
2.2 KEY CONSIDERATIONS FOR DESIGNING AND USING ASSESSMENTS IN LANGUAGE LEARNING ENVIRONMENTS 6

3 RIGHTS AND RESPONSIBILITIES OF TEST TAKERS 7

3.1 TEST TAKER RIGHTS 7
3.2 TEST TAKER RESPONSIBILITIES 8

4 GUIDELINES FOR DESIGNING AND USING FORMAL LANGUAGE TESTS 9

4.1 VALIDITY 9
4.2 COMMUNICATION & REPORTING 10
4.3 CONSISTENCY 11
4.4 FAIRNESS, INCLUSION & ACCESSIBILITY 12
4.5 SECURITY & SAFETY 12
4.6 EQUIVALENCE 13

5 GUIDELINES FOR DESIGNING AND USING ASSESSMENT IN LANGUAGE LEARNING ENVIRONMENTS 14

5.1 SUITABILITY & USE 14
5.2 COMMUNICATION & REPORTING 15
5.3 FAIRNESS, INCLUSION & CONSISTENCY 15

6 GLOSSARY 17

7 DEVELOPMENT OF THE GUIDELINES AND ACKNOWLEDGEMENTS 21

1 Overview of the ILTA Guidelines

Language tests/assessments[1] are used throughout the world for decisions that affect the lives of millions of people. Therefore, it is important that people who make language assessments ("designers" or "developers"), people who provide or sell language tests ("test providers"), people who use language assessments for particular decisions ("test/score users"), and people who take language tests ("test takers") have a common understanding of best practice in language assessment. These Guidelines set out basic information about best practice in language assessment to promote fair, valid and transparent uses of language assessments.

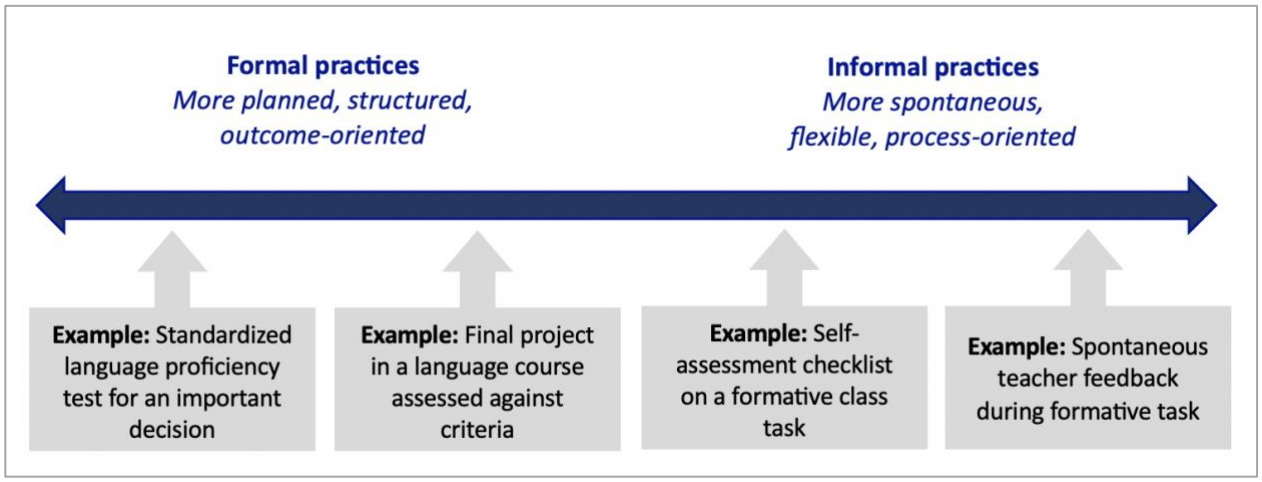
1.1 Who are the Guidelines for?

These Guidelines are for anyone who designs, provides, uses, evaluates or undergoes language assessment, including test designers, test providers, policy makers, teachers and people who are assessed.

1.2 What is covered in the Guidelines?

In the Guidelines, assessment is conceived of as a range of practices, from those that are more formal, planned, and structured to those that are more informal, spontaneous, and flexible (see Figure 1). The Guidelines cover a range of assessment practices, but they focus mainly on those that carry some weight or consequences for assessees. These are practices that tend to be more structured, outcome-oriented and planned.

Figure 1: Range of assessment practices



[1] "Assessment" is used as an umbrella term which includes formal testing and other assessment practices throughout the Guidelines.

1.3 How to use the Guidelines

There are two main sets of guidelines: **Guidelines for designing and using formal language tests** (detailed in Section 4) and **Guidelines for designing and using assessment in language learning environments** (detailed in Section 5). The **Summary of key considerations** provided in Section 2 is intended as a brief overview of important points covered across the full range of assessment practices. The summary should be used as a quick reference only, and not in place of the more detailed guidelines in Sections 4 and 5. Section 3 sets out the **rights and responsibilities** of test takers (and their legal guardians where appropriate). Test providers and test users must ensure that test takers are given these rights. The rights and responsibilities relate mainly to high-stakes standardized tests, though some aspects are relevant to any kind of formal assessment. A **glossary** is included (Section 6) which clarifies terms as they are used in the Guidelines.

The first set of Guidelines focuses on **designing and using formal language tests** (Section 4). This section provides guidelines for formal testing situations, such as large-scale standardized tests. Standardized test providers and users will be mainly concerned with Section 4. Teachers who are responsible for preparing students for standardized tests will find the formal test guidelines relevant to their work. Educators who design formal tests will also find Section 4 useful but will not need to adhere to points explicitly directed at large-scale standardized tests.

The second set of Guidelines (Section 5) focuses on **designing and using assessment in language learning environments,** also known as "classroom-based assessment". This section is relevant to assessment that occurs in a designed sequence of learning for a particular group of learners, including online learning communities. Section 5 focuses on various types of formal and informal assessment used in language learning contexts. It includes **formative assessments,** which provide information about what to consider in future instruction and learning, and **summative assessments,** which provide a summary of prior learning (see Glossary for more detailed definitions). However, the terms 'formative' or 'summative' are descriptions of assessment purposes rather than assessment methods.

2 Summary of key considerations

Key points from the range of practices represented in the Guidelines are set out below. Detailed guidelines relating to each of these major themes and key ideas can be found in Sections 4 and 5.

2.1 Key considerations in designing and using formal language tests

Validity

A test must be demonstrably suited to its stated purpose, the population and the context in which it is used so that the test scores are trustworthy and meaningfully related to the decisions and judgements they inform.

Communication & reporting

Clear and accessible information must be provided about the intended test purpose/use, the intended test taker population, what the test is supposed to measure (the "construct"), the test format, the scoring methods and the delivery mode.

Scores should be presented in such a way that they can be easily accessed and fully understood by test users and test takers.

High-stakes standardized tests should provide regular, timely, publicly-available technical reports for test users.

Consistency

Test scores on a particular test must be consistent and comparable across time, settings, delivery variations, versions (forms) of the test and intended populations.

Tests should be administered and scored/rated by suitably qualified and trained personnel and/or reliable systems using established and equitable procedures as appropriate to the test purpose, the tested population and the testing context.

Security & safety

Testing environments (physical facilities or technological environments), test delivery personnel and test content/materials should be managed in such a way that no test taker is able to gain an unfair advantage, or is unfairly penalised.

Fairness, inclusion & accessibility

Test takers must be treated equitably and fairly. Test takers' rights should be respected.

Test design and use should be ethical. Designers, providers and users should strive to maximise positive impact, while preventing potential negative impact.

Equivalence

Evidence is required to support any claims that different test scores, different test forms and different test modes are equivalent.

2.2 Key considerations for designing and using assessments in language learning environments

Suitability & Use

Assessment (including questions, activities, tasks and criteria) should be designed to target an explicitly stated area of knowledge, skill or ability at an appropriate level of challenge for the particular cohort of learners.

Assessment should inform future learning and instruction (formative tasks) and/or accurately represent progress over time (summative tasks).

Communication & Reporting

Clear, timely information about assessment tasks and evaluation methods should be provided to learners in ways they can understand.

Results/feedback should be communicated to learners in ways they can understand, with consideration of the characteristics and needs of the learners, and the demands/purpose of the assessment task.

Fairness, inclusion & consistency

All learners should have appropriate and equitable opportunities to participate in assessment activities to the best of their abilities.

Assessment should be marked consistently and in line with any designated standards.

3 Rights and responsibilities of test takers[2]

3.1 Test taker rights

Test providers and test users must ensure that test takers are given the following rights. All test takers (and their legal guardians where appropriate) have the right to:

a) be informed of their rights and responsibilities;

b) be treated with respect and impartiality, regardless of age, disability, ethnicity, gender, national origin, religion, sexual orientation or other personal characteristics;

c) receive an explanation prior to testing about the purpose(s) of the test, the format of the test, who will receive their scores, and the planned use(s) of the scores;

d) ask questions about the test prior to testing;

e) complain about an assessment or appeal a score in a timely, respectful way and be informed of procedures for complaints or appeals prior to being assessed;

f) appeal their score in a process that allows a fresh and fair assessment of their responses, for example, a different human rater, or, in the case of fully automated assessments, a human rater;

g) access sufficient information about the meaning of their score and its limitations, for example, the degree of measurement error;

h) not be penalised for appealing a score;

i) be tested with measures that meet professional standards and that are appropriate, given the use of the scores;

j) be informed before being tested about any testing accommodations (special arrangements) for test takers with a disability;

k) know in advance of testing when the test will be administered, if and when test results will be available and if there is a fee for testing services that test takers are expected to pay;

l) have tests administered and test results interpreted by appropriately trained individuals who follow professional codes of ethics;

m) be informed of whether a test is optional and, if so, the consequences of taking or not taking the test;

n) receive an explanation of test results within a reasonable timeframe and in commonly understood terms;

o) be informed of any technology required to take the test and be provided ample time to become familiar with the technology and the functionalities of the system;

[2] This section on the rights and responsibilities of test takers is adapted from the *Rights and responsibilities of test takers: Guidelines and expectations* published by the American Psychological Association (1998).

p) know how and to what extent technology (e.g., AI, automated scoring) is involved in the development, scoring, and administration of tests and assessments;

q) be informed about test security measures and the storage and use of their individual data;

r) be assessed on complex language constructs for high-stakes purposes using methods which involve human judgement (i.e. not solely through automated methods);

s) have human oversight in high-stakes decisions about them.

3.2 Test taker responsibilities

All test takers (or legal guardian, as appropriate) have a responsibility to:

a) treat others with respect during the testing process, and do nothing to prevent other test takers from performing to the best of their abilities;

b) access the available information about the delivery/administration of the test in advance of testing;

c) inform testing personnel in advance of testing if they require a testing accommodation;

d) maintain the integrity of the test by abiding by the stated test conditions and not interfering with the capacity of other test takers to abide by the same conditions;

e) maintain the integrity of the test by not engaging in malpractice (e.g., cheating or fraud);

f) inform appropriate person(s) responsible for testing through appropriate channels and in a timely manner if something unfair about the test has affected their results.

4 Guidelines for designing and using formal language tests

4.1 Validity

4.1.1 A test must be demonstrably suited to its stated purpose, the population and the context in which it is used so that the test scores are trustworthy and meaningfully related to the decisions and judgements they inform.

4.1.2 Test design and development should be driven by theory and empirical evidence about language use in the target domain so that the language, tasks, skills, abilities or knowledge sampled in the test have demonstrated relevance and significance in the target domain.

4.1.3 Tests should sample enough of the relevant language, tasks, skills, abilities or knowledge that reliable and trustworthy inferences can be made based on test takers' performances.

4.1.4 Tests should minimise interference from sources that are not relevant to the test construct, for example, irrelevant tasks, unclear instructions, unfamiliar format/functionality, noisy environments.

4.1.5 Any human or technological processes or agents involved in the development and administration of tests (e.g., item writing/generation, human/automated interlocutors, automatic speech recognition, human/automated scoring) must be capable of accuracy, domain-relevance and consistency in processing naturalistic language use. Capability should be demonstrated through regular monitoring, training/development and validation research.

4.1.6 Humans should be involved in high-stakes judgements about language abilities where complex language constructs are assessed.

4.1.7 Test providers and users should seek to understand the impacts of their tests and test methods, including the impact on learning and teaching and test preparation practices ("test washback"). Any actual or potential negative impacts of tests that stem from known and controllable aspects of test design or use should be anticipated to the extent possible and addressed. Any aspects of test design and the impact of its use that are deemed uncontrollable or not controllable to an acceptable level should be disclosed, monitored and addressed to the extent possible.

4.1.8 Score interpretations that are based on the performance of a particular population (i.e., "norm-referenced testing") must be based on evidence from a population that is comparable to the intended test taker population and appropriate in relation to the intended score use.

4.1.9 If test performances and scores are linked to standards or another kind of level description (i.e., "criterion-referenced testing"), the levels/descriptions and their relationship to test scores must be verified using a range of appropriate methods such as expert judgement, and not limited to correlational studies.

4.1.10 Validity evidence for high-stakes uses of tests should be sought on an on-going basis and include studies conducted by independent researchers. If a test undergoes changes in methods or design revisions, validity evidence that it remains suitable for use should be available prior to live test administration.

4.1.11 When promoting or using tests for significant decisions, test providers and users should consider the limits of standardized tests as one-off samples of language use.

4.2 Communication & reporting

4.2.1 **Clear and accessible information must be provided about the intended test purpose/use, the intended test taker population, what the test is supposed to measure (the "construct"), the test format, the scoring methods and the delivery mode**.

4.2.2 Tests should be described in detail in language that is understandable to intended test users. Information should be available about all aspects of tests including test conditions, test tasks, scoring methods, weightings towards overall scores (e.g., of tasks, of criteria and of scored features) and any other aspects of a test that is relevant to test taker performance and scoring.

4.2.3 For standardized tests, the test construct must be clearly defined with an explanation of how the construct and sub-constructs are operationalized in each component of the test.

4.2.4 Information about the test for users and test takers should be transparent regarding the use of technology in test development, delivery, scoring, and administration; for example, the use of automated scoring systems, the use of Artificial Intelligence (AI) in delivery, development or scoring, and the use of automated systems to detect or prevent cheating or malpractice.

4.2.5 Materials, technology and resources (e.g., AI tools, translation software, dictionaries) that are required, allowed or prohibited while doing an assessment should be clearly communicated.

4.2.6 Opportunities to become familiar with all aspects of a test and any technological functionalities should be freely available to test takers prior to being tested.

4.2.7 Information that is important for test takers to understand in order to complete the test (e.g., timing, venue, technological requirements, task format and instructions) should be provided in such a way that the intended test takers can access and understand.

4.2.8 The technological requirements for administering the test must be clearly communicated.

4.2.9 Information about available accommodations should be accessible and clearly communicated.

4.2.10 Information about appeals processes should be accessible and clearly communicated, including any constraints on appealing.

4.2.11 Test security requirements and procedures should be communicated to test takers prior to testing.

4.2.12 Test information, including information about any changes to test design, should be made accessible to test takers and test users so that there is time for users to make informed decisions about test use, and time for test takers to prepare for tests.

4.2.13 Test designers, providers and users must not make or endorse false or misleading claims about the test (e.g., claims about possible uses that are not supported by validity evidence, or claims about endorsements that have not been granted). Any known misuses of tests should be addressed.

4.2.14 Scores should be presented in such a way that they can be easily accessed and fully understood by test users and test takers.

4.2.15 Score reports should contain enough information so that test takers and other stakeholders (e.g., teachers, parents, professional registration bodies, educational institutions) can draw appropriate and meaningful inferences from them.

4.2.16 Score reports should be delivered according to a published/agreed schedule in a reasonable timeframe.

4.2.17 High-stakes standardized tests should provide regular, timely, publicly-available technical reports for test users.

4.2.18 Technical reporting should be written in accessible language that intended users can understand, with key concepts defined. The basis for score calculations should be explicit. Regular reporting should include:

i. the performance of test items and tasks;
ii. appropriate reliability estimates, with clear descriptions of what test data they represent;
iii. error estimates for the test and sub-tests, so that users understand the limits of the scoring at important score thresholds.
iv. bias investigations into the performance of sub-populations;
v. investigations into potentially variable aspects of the test.

4.2.19 Validation research should be made freely accessible. Test providers should support open-access publication, and are encouraged to follow open science principles.

4.3 Consistency

4.3.1 Test scores on a particular test must be consistent and comparable across time, settings, delivery variations, versions (forms) of the test and intended populations.

4.3.2 Tests should be administered and scored/rated by suitably qualified and trained personnel and/or reliable systems using established and equitable procedures as appropriate to the test purpose, the tested population and the testing context.

4.3.3 Reliability should be investigated and reported for any aspects of a test where consistency may be threatened, such as the involvement of different raters (whether for human scoring or calibrating automated scoring mechanisms), and the use of different forms of a test.

4.3.4 A test should capture enough information about test constructs and sub-constructs from each test taker to support adequate reliability.

4.3.5 Tasks, texts and items should be well constructed with clear relevance to the test construct and domain of use.

4.3.6 Items and tasks in high-stakes assessments should be reviewed and trialled prior to use in a live test administration and any weak components should be revised or removed.

4.3.7 Reliability monitoring should inform ongoing procedures such as rater training, item writer training and algorithm calibration.

4.3.8 Score processing must be routinely monitored for errors.

4.3.9 Test providers should keep their testing system functional without undue interruptions. Technical support should be available to test takers and plans should be in place to respond to technical difficulties.

4.4 Fairness, inclusion & accessibility

4.4.1 Test takers must be treated equitably and fairly. Test takers' rights should be respected.

4.4.2 Test design and use should be ethical. Designers, providers and users should strive to maximise positive impact, while preventing potential negative impact.

4.4.3 Where possible, test taker perspectives should be considered in the design, development and use of tests.

4.4.4 Large-scale standardized test providers should have a published fairness, inclusion, and accessibility policy which covers test development, delivery and administration.

4.4.5 Appropriate accommodations should be offered so that individual test takers with disabilities experience the test equitably.

4.4.6 The delivery mode of a test should not prevent individuals from performing to the best of their ability in relation to the test construct.

4.4.7 For high-stakes testing, bias investigations for different sub-populations (e.g., gender bias, first language bias) should be part of regular monitoring and reporting for test users. Action should be taken to remove sources of bias if any are detected, and to mitigate any bias found to significantly impact results.

4.4.8 Tests that use automated methods of item generation or scoring should ensure that biases in source data (e.g., Large Language Models) are not replicated in the test methods through careful selection of training data, regular monitoring of test content and processes, and calibration with human raters.

4.4.9 For high-stakes testing, clear and transparent complaint and appeal processes should be in place. Score appeals should allow a fresh and fair scrutiny of scores, for example, through the use of a different rater (for human rating) or a human rater (for automated scoring).

4.5 Security & safety

4.5.1 Testing environments (physical facilities or technological environments), test delivery personnel and test content/materials should be managed in such a way that no test taker is able to gain an unfair advantage, or is unfairly penalised.

4.5.2 Test environments, whether they are physical facilities or digital environments, should enable effective invigilation and oversight and provide a safe and secure environment for test takers and test administrators.

4.5.3 Technologies involved in test security (e.g., remote proctoring, plagiarism detection, voice identification) should be demonstrably robust, reliable and fit-for-purpose, with evidence provided for their effectiveness and appropriateness. Test security technologies and processes should be commensurate with the test purpose. Test security measures should be balanced with concern for test takers' privacy.

4.5.4 Score reporting methods should ensure privacy and confidentiality. Score reports should be made available only to those who are known by test takers as official or predesignated recipients.

4.5.5 Collection and storage procedures for test data (e.g., personal information, test responses, test scores) should ensure security, confidentiality and privacy.

4.5.6 If an event or action occurs that calls into question the integrity of the test, the problem should be identified and addressed. Any remedial action to be taken to offset the negative impact on the affected test takers should be promptly announced.

4.6 Equivalence

4.6.1 Evidence is required to support any claims that different test scores, different test forms and different test modes are equivalent.

4.6.2 For tests that are available in different delivery modes for the same purpose (e.g., a 'pencil and paper' and a computer-delivered form of the test), evidence that there is no difference in outcomes for test takers is required.

4.6.3 Equivalence between scores on different tests should be established empirically using a sufficiently robust sample across the range of score levels and a suitable concordance method which enables a close comparison of the tests.

4.6.4 Equivalence between scores on different tests should only be sought and claimed between tests with demonstrably similar constructs and intended uses.

4.6.5 Equivalence claims between test scores on different tests should be presented in language users can understand and in a way that clearly sets out the limits of the score relationship.

5 Guidelines for designing and using assessment in language learning environments

This section contains guidelines that are relevant to instructional contexts in which educators design, administer and mark assessment for the purposes of informing and promoting learning (known as "formative assessment") and for the purposes of summarizing achievement following a period of learning (known as "summative assessment"). Educators will find relevant information about formal tests they may design or use in Section 4. This section focuses on other types of formal and informal assessment that are used in language learning contexts, including online environments.

5.1 Suitability & use

5.1.1 **Assessments (including questions, activities, tasks and criteria) should be designed to target an explicitly stated area of knowledge, skill or ability at an appropriate level of challenge for the particular cohort of learners.**

5.1.2 Assessment tasks and evaluation criteria should be designed in alignment with the necessary curriculum outcomes and standards and the needs/level of the learners.

5.1.3 Prior to undergoing a summative assessment, learners should be given an opportunity to learn the knowledge or develop the skills or abilities that are to be assessed.

5.1.4 When designing to a set standard for a particular cohort, a range of difficulty levels should be incorporated in the assessment task through, for example, including components that are relatively easy or more scaffolded as well as components that are more challenging or more independently achieved.

5.1.5 Tasks should be carefully reviewed prior to using them with learners, and after each use. If a problem is found during the use of an assessment task, efforts to mitigate the problem should be made, for example, through adjustments to marks.

5.1.6 Educators who are responsible for assessment design must have appropriate expertise and be familiar with the context of assessment.

5.1.7 Educators and educational institutions must monitor, and be accountable for, the use of automation in assessment procedures and decisions (e.g., AI-generated tasks, automated feedback or scoring). Significant judgements about learners' achievement, particularly judgements about complex language constructs, should involve human judgement and not be based solely on automated methods.

5.1.8 Assessment should inform future learning and instruction (formative tasks) and/or accurately represent progress over time (summative tasks).

5.1.9 Formative assessment should generate some kind of feedback for learners, whether it is discursive (e.g., an error code, a reflection or comment), interactional (e.g., a discussion) or categorical (e.g., a score).

5.1.10 Marks and/or feedback should be returned in a reasonable timeframe given the nature of the task, the nature of the feedback and the sequencing of assessments.

5.1.11 Opportunities for learners to review their tasks/outcomes or act on feedback should be provided, so learners are enabled to take responsibility for their own learning and understand how to move forward using the feedback.

5.1.12 The constraints of human/automated feedback should be clearly explained to learners (e.g., which kinds of error are noted or which aspects of a task a feedback mechanism can and cannot address).

5.1.13 Learners should be given designated opportunities to ask questions about the marks/feedback they receive.

5.1.14 Peer assessment tasks should be well guided by the teacher in terms of identifying the appropriate focus, using the method of assessment and interacting respectfully.

5.2 Communication & reporting

5.2.1 Clear, timely information about assessment tasks and evaluation methods should be provided to learners in ways they can understand.

5.2.2 Prior to an assessment, learners should be informed about the following aspects, as appropriate for their age and the type of assessment:

i. what is being assessed;
ii. the nature of the assessment tasks and conditions;
iii. how assessments are marked and how results and any feedback is generated (e.g., automated or human);
iv. how the assessment fits within the broader learning sequence and learning outcomes including which assessment tasks will count towards final grades;
v. how assessment tasks are weighted in calculating their final results;
vi. what resources and tools they are allowed to use in completing an assessment;
vii. for group tasks, how individual scores and group scores will be allocated and used;
viii. who can know their result.

5.2.3 Learners should be given opportunities to ask questions about assessment tasks.

5.2.4 Results/feedback should be communicated to learners in ways they can understand, with consideration of the characteristics and needs of the learners, and the demands/purpose of the assessment task.

5.2.5 Learners should have access to their results.

5.2.6 Where possible, feedback should communicate information about future learning.

5.2.7 Results/feedback should be communicated sensitively so that there is no detrimental impact on learner well-being.

5.2.8 Assessment data must be stored securely and kept confidential. Results should only be shared in the course of professional duties (e.g., with parents, educational authorities, other teachers).

5.3 Fairness, inclusion & consistency

5.3.1 All learners should have appropriate and equitable opportunities to participate in assessment activities to the best of their abilities.

5.3.2 All learners should be able to complete an assessment task with the same conditions (e.g., time allowed, venue, access to resources/technology, word limit, access to preparation opportunities) unless an accommodation is in place.

5.3.3 Appropriate accommodations should be offered for learners with disabilities.

5.3.4 Interference from sources that are not relevant to the assessment focus should be minimized wherever possible, for example, an inadequate physical venue, a distracting environment or problems with technology.

5.3.5 Learners should be familiar with the assessment mode (e.g., using a keyboard or digital functionality), or they should be given sufficient time to become familiar with the assessment delivery mode.

5.3.6 Where there are significant differences in the characteristics of a class or cohort (e.g., different language/literacy backgrounds), explicit differentiation should be made in relation to task design and/or marking criteria and standards so that learners are able to access achievement at their levels.

5.3.7 Where possible, learner perspectives should be included in the assessment design process, for example, seeking feedback on appropriate timing or the extent to which tasks are sufficiently engaging.

5.3.8 Assessment should be marked consistently and in line with any designated standards.

5.3.9 If multiple educators are using same assessment across different cohorts, common understandings amongst the educators about tasks, marking methods and standards should be established in advance.

5.3.10 For assessments that have higher stakes (e.g., summative assessments), moderation processes such as consensus marking or double-marking should be used.

5.3.11 Efforts should be made to ensure that no individual gets an unfair advantage through cheating or other deceptive practices.

5.3.12 Learners should be given the chance to appeal the result of a summative assessment if they consider it to be unfair.

6 Glossary

assessment

"Assessment" is used as an umbrella term to refer to an evaluation instrument and procedure that involves collecting, evaluating, and interpreting evidence to make informed judgments about a specified attribute, ability, skill or area of knowledge. An "assessment" is also used to refer to a specific evaluation practice, especially one that is not highly constrained in timing and access to resources. (See "test" and 'formal testing/assessment").

Artificial Intelligence (AI)

In the Guidelines, "artificial intelligence" refers to the use of computer technology that attempts to mimic or replace a human's action, thoughts, or appearance.

automated or AI-assisted tests

"Automated" or "AI-assisted" tests are tests that utilize artificial intelligence or other machine-based methods to assist in assessment processes such as item generation, adaptive testing, scoring, and providing personalized feedback.

classroom-based assessment

"Classroom-based assessment" refers to the evaluation of learners' language abilities in an instructional setting to provide feedback and inform instructional plans.

construct

A "construct" is what an assessment is designed to measure. "Sub-constructs" are components of more general constructs. For example, a test of speaking ability might comprise sub-constructs such as "range of vocabulary" or "interactional ability".

criterion-referenced testing

"Criterion-referenced testing" involves evaluating a test taker's performance against predetermined criteria or standards and providing information on whether the test taker has met the set requirements.

formal testing/assessment

"Formal testing" or "formal assessment" refers to planned, structured instruments or tasks with controlled conditions. Formal tests are typically used for summative and/or high-stakes purposes such as to measure achievement at the end of a course of learning or to measure language proficiency for professional registration. The most uniform and controlled type of formal assessment is standardized testing.

formative assessment

"Formative assessment" is a process-oriented evaluation conducted during the learning process to provide feedback which could lead to improvement in language teaching and learning. Formative assessment has a complementary relationship with summative assessment in a course of learning. (See "Summative assessment".)

high-stakes tests

"High-stakes tests" have outcomes which carry very significant individual or societal consequences, such as determining an individual's citizenship status or determining education funding.

impact

The "impact" of an assessment is a broad term which covers any effect or consequence of the use of an assessment. Impact includes "washback", the effect of an assessment on learning and teaching, as well as its broader effects on individuals, groups and socio-educational systems.

informal assessment

"Informal assessment" refers to spontaneous, unstructured, and often ongoing evaluation methods such as observations, checklists, peer- and self-assessments and ad hoc feedback which are aimed at providing feedback for improvement rather than ranking or scoring performances for summative or official purposes.

norm-referenced testing

"Norm-referenced testing" involves evaluating a test taker's performance by comparing it to the performance of a group, allowing for an understanding of the test taker's ability in relation to their peers.

reliability

"Reliability" refers to the consistency of the results of a particular test, to what extent they are generalizable and therefore comparable across time, test forms and settings.

scores

The term "scores" (also "results" or "marks" in the Guidelines) refers to any categorization based on test responses/performances, whether it is a numerical value or some other categorization system such as levels, rankings, grades or category descriptions.

summative assessment

"Summative assessment" is an assessment or sequence of assessments that are used as a basis for reporting learners' achievements over a period of learning. A summative assessment may take many forms, from group projects to standardized tests. Summative assessment has a complementary relationship with formative assessment in a course of learning. (See "formative assessment".)

target domain

The "target domain" of the test is the language of the area of life, experience or learning that the test is aimed at assessing. The following are examples of target domains:

- The language of a professional domain, such as the language of nursing
- The curriculum content and learning outcomes of a period of learning such as a language studied in a semester of secondary school

technology-mediated tests

"Technology-mediated tests" are assessments that are delivered or facilitated through technological means including computer-based testing, online exams, or the use of digital tools for assessment purposes.

test

A "test", also known as an "exam" or "examination", is a formal procedure that is usually constrained in timing and available resources. A test is designed to measure a specified attribute, ability, skill or area of knowledge. It is a specific kind of assessment. (See "assessment")

test designer/test developer

A "test designer", also known as a "test developer", is an individual, institution, or organization that is responsible for designing and constructing formal language tests including test specifications, tasks, criteria, and rubrics. The term "test design" is sometimes used in relation to the initial stages of test construction, including the test specifications; "test development" often relates to ongoing test construction, for example, for new versions of a standardized test.

test provider

A "test provider", also known as a "testing agency", is an institution or company that is responsible for producing and validating tests for specified uses. Test providers often also administer their tests.

test specifications

"Test specifications" are detailed guidelines outlining the content, format, structure, and other relevant information of a language test, providing a blueprint for the development and administration of the test.

test taker

A "test taker", also known as an "examinee", "candidate", "testee" or "assessee", is a person who undergoes a test, an examination, or an assessment to demonstrate their language abilities. In learning environments, "students" and "learners" also undergo assessment.

test user

A "test user", also "score user", is a person or entity who uses a test for a particular purpose in a specific context. Test users usually select an existing test provided by another entity, rather than design it themselves. The following are examples of "test users":

- A government that selects and implements a language test provided by a testing company as part of a national migration policy
- An institution that requires a certain score on a selected (externally-developed) language test as an entrance requirement
- A teacher who selects a published test for use with a group of students

test writer

A “test writer”, also known as an “item writer”, is an individual responsible for constructing test materials including items, tasks, and texts.

validity

“Validity” refers to the accuracy, suitability and trustworthiness of test scores. Validity is judged in relation to how scores are interpreted or used. “Validation studies” are carried out to determine the strength of the relationship between the test methods and the ways scores are interpreted or used. The interpretation of a test score can be valid only if the test offers an accurate and genuine representation of the knowledge, skill or ability it is supposed to measure.

Validity: Example 1 – Language proficiency testing for professional purposes: A particular test score on a test of Aviation English is understood by authorities to indicate that the test taker (a trained pilot) can communicate to a level that ensures safety during flights. For this understanding to be accurate, the test must have sampled and reliably scored sufficient amounts of relevant and genuine language use by the pilot. A validation study might aim to find out if the score does in fact represent a safe level of communication in the cockpit. Positive evidence from various validation studies allows us to trust the claim that the pilot can communicate safely.

Validity: Example 2 – Classroom-based assessment: A mark of above 70 out of 100 on a final class project is described to students and their guardians as meeting certain national curriculum outcomes for learning an additional language. For this message to be accurate, the task must have sampled (and consistently marked) adequate, genuine evidence of the student’s ability to perform the specific outcomes described in the curriculum. A validation study might aim to find out if the mark on this project (or on a sequence of relevant assessments across the course) actually demonstrates that the student can perform the language functions described in the curriculum outcomes.

7 Development of the Guidelines and acknowledgements

The first version of the ILTA Guidelines for Practice was presented as a draft at the 2005 ILTA meeting in Ottawa and then circulated among members for further consideration. The Guidelines were adopted at the 2007 ILTA meeting in Barcelona and were reviewed and revised between 2018 and 2020. The Guidelines were reviewed and revised again in 2023-2024, with a view to increasing their accessibility and breadth, including the development of technology. The 2024 revision was first drafted by a working group and then circulated for feedback from the ILTA membership, after which further revisions were made. In the process of the 2024 revision, the structure of the guidelines was modified substantially to enhance coherence and promote wider engagement.

Those involved in the original drafting were Charles Alderson, Alan Davies (chair), Glenn Fulcher, Liz Hamp-Lyons, Antony Kunnan, Charles Stansfield and Randy Thrasher. The 2018-2020 revisions were drafted by Vivien Berry, Benjamin Kremmel (chair) and India Plough. The 2024 revisions were drafted by Dan Douglas (chair), Luke Harding, Yan Jin, Benjamin Kremmel, Susy Macqueen and Erik Voss. The section on the rights and responsibilities of test takers is adapted from the *Rights and responsibilities of test takers: Guidelines and expectations* published by the American Psychological Association (1998).