Cir 318
AN/180

Language Testing Criteria for Global Harmonization

Approved by the Secretary General
and published under his authority

International Civil Aviation Organization

Cir 318
AN/180

Language Testing Criteria for Global Harmonization

Approved by the Secretary General and published under his authority

International Civil Aviation Organization

Published in separate English, Arabic, Chinese, French, Russian and Spanish editions
by the INTERNATIONAL CIVIL AVIATION ORGANIZATION
999 University Street, Montréal, Quebec, Canada H3C 5H7

For ordering information and for a complete listing of sales agents
and booksellers, please go to the ICAO website at www.icao.int

ICAO Cir 318, *Language Testing Criteria for Global Harmonization*
Order Number: CIR318
ISBN 978-92-9231-271-8

FOREWORD

1. Assembly Resolution A36-11*, Proficiency in the English language used for radiotelephony communications,* directed the Council to support Contracting States in their implementation of the language proficiency requirements by establishing globally harmonized language testing criteria.

2. Organizing aviation language testing is one of several steps necessary to effectively implement the ICAO language proficiency requirements. ICAO has previously published detailed “Implementation Guidelines” for aviation language requirements; these are available at http://www.icao.int/fsix/lp.cfm.

3. While the *Manual on the Implementation of ICAO Language Proficiency Requirements* (Doc 9835), published in September 2004, provided some guidance on testing, users of the manual — including licensing authorities, air operators, air navigation service providers, and language training and testing services — have indicated that more detailed guidance on language testing is needed to effectively implement the language proficiency requirements. The purpose of this circular is to address that need.

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

1. OVERVIEW

1.1 Purpose

The purpose of this circular is to provide guidance to civil aviation authorities and test service providers in processes for testing candidates in accordance with the ICAO language proficiency requirements. In particular, it provides recommended criteria to guide the development or selection of aviation language testing programmes, as well as additional guidance material in that regard. The recommended aviation language testing criteria outlined in this circular were drawn from principles of best practices by the Proficiency Requirements in Common English Study Group (PRICESG) in 2005. They are intended to support the harmonization of global aviation language testing.

1.2 Target Audience

1.2.1 This publication will be useful to civil aviation and licensing authorities that oversee language testing and will provide State authorities, airlines and air navigation service providers with a set of practical tools. Civil aviation and licensing authorities may use these criteria:

a) as a guide to development, should they decide to assign national resources to develop aviation language testing; and

b) as a checklist against which to compare and assess externally developed aviation language tests;

1.2.2 Language testing organizations may use the criteria as a guide to provide the information and evidence to document that they comply with the criteria and to establish the integrity of their test.

1.3 Applicability to all languages

The ICAO language proficiency requirements (LPRs), which apply to **all** languages used in international radiotelephony communications, create a significant testing requirement. This is particularly true with respect to English which is the language for which most training and testing programmes need to be developed. While this circular focuses on criteria guiding the development or selection of language tests in English, the principles apply equally to tests developed for any language used for international radiotelephony communications.

1.4 Context

The recommended criteria in this guidance material are considered appropriate to the diverse contexts in which aviation language testing occurs. The principles underlying these criteria suit various operational and regulatory needs at various points of application within each particular administration.

1.5 Structure

Chapter 1 – *Introduction.* Presents the background and context, with references to other appropriate ICAO publications and guidance material.

Chapter 2 – *Recommended Criteria for Aviation Language Testing.* Presents the criteria. Each criterion includes: what it means, why it is important and, where applicable, additional information.

Chapter 3 – *Checklist.* Provides the recommended criteria in checklist format.

Glossary of Language Proficiency and Language Testing Terms and Acronyms

2. BACKGROUND

2.1 ICAO language proficiency requirements

2.1.1 The decision to address language proficiency for pilots and air traffic controllers was first made by the 32nd Session of the Assembly in September 1998 as a direct response to several fatal accidents, including one that cost the lives of 349 persons, as well as to previous fatal accidents in which the lack of proficiency in English was identified as a contributing factor.

2.1.2 Subsequently, the Air Navigation Commission initiated the development of language provisions in the following Annexes to the Convention on International Civil Aviation and PANS:

a) Annex 1 — Personnel Licensing;

b) Annex 6 — Operation of Aircraft;

c) Annex 10 — Aeronautical Telecommunications, Volume II; and

d) Annex 11 — Air Traffic Services

e) PANS-ATM

2.1.3 In March 2003, the ICAO Council adopted a comprehensive set of Standards and Recommended Practices (SARPs) that strengthened language proficiency requirements for pilots and air traffic controllers involved in international operations. These language proficiency requirements affirmed that ICAO standardized phraseology should be used whenever possible and required that when phraseology is not applicable, pilots and air traffic controllers should demonstrate a minimum level of proficiency in plain language. The effective use of plain language is vital in routine operational situations in which phraseology provides no “ready-made” form of communication and is especially critical in unusual or emergency situations. The minimum skill level requirements are embodied in the ICAO language proficiency rating scale and the holistic descriptors that appear in Attachment A and Appendix 1 of Annex 1, respectively.

2.1.4 As of 5 March 2008, the ability to speak and understand the language used for radiotelephony that is currently required for pilots, air traffic controllers and aeronautical station operators should be demonstrated based on the holistic descriptors and language proficiency rating scale to at least Level 4. Level 4 is considered the minimum level of proficiency to ensure an acceptable level of safety. Additionally, since November 2003, Annex 10, Volume II, has required the availability of English at all stations on the ground serving designated airports and routes used by international air services.

2.2 ICAO support and guidance

ICAO has provided Contracting States with implementation support in a number of ways, primarily through symposia, regional seminars and workshops, and the publication of guidance material:

a) First ICAO Aviation Language Symposium — September 2004, Montréal;

b) Doc 9835, *Manual on the Implementation of ICAO Language Proficiency Requirements,* published in September 2004;

c) Regional seminars — Japan (2004); Argentina, Azerbaijan, Belgium, Ukraine (2005); China (Hong Kong), France, Mexico, Senegal (2006); Egypt, France, Germany (2007);

d) Rated Speech Samples CD — A set of Rated Speech Samples developed and published by ICAO in 2006. Samples of speech at ICAO Pre-Operational Level 3, Operational Level 4, and Extended Level 5 are provided, with full explanations and rationale for the assignment of each level. This CD is available for purchase from ICAO (contact: sales@icao.int);

e) Second ICAO Aviation Language Symposium — May 2007, Montréal;

f) Regional workshops on the development of States' implementation plans for language proficiency requirements — Belarus, Peru (2007); El Salvador, Mexico, Senegal, Thailand, Uganda, United Arab Emirates (2008);

g) Implementation of language proficiency requirements website (http://www.icao.int/fsix/lp.cfm) — Further to the adoption of Assembly Resolution A36-11, States can find information concerning States' level of compliance with the language proficiency requirements and their implementation plans, as well as other implementation guidance, on this website; and

h) ICAO Frequently Asked Questions (FAQs) website (http://www.icao.int/icao/en/trivia/ peltrgFAQ.htm) — updated with information regarding the ICAO language proficiency requirements.

2.3 ICAO Assembly Resolution A36-11

Assembly Resolution A36-11, *Proficiency in the English language used for radiotelephony communications,* directed the Council to support Contracting States in their implementation of the language proficiency requirements by establishing globally harmonized language testing criteria. The circular is a direct outcome of A36-11.

2.4 Frame of reference

2.4.1 While some regional and national language testing certification programmes exist and some testing programmes are self-regulated, no universal system of aviation language testing certification has yet been developed.

2.4.2 The recommended language testing criteria presented herein are aligned with other ICAO guidance material. This circular should be considered a supplement to and not a replacement for other ICAO publications. Users should be familiar with the following:

a) Doc 9835, *Manual on the Implementation of ICAO Language Proficiency Requirements*;

b) ICAO Language Proficiency Requirements Rated Speech Samples; and

c) the Frequently Asked Questions page at http://www.icao.int/icao/en/trivia/peltrgFAQ.htm.

2.5 Scope

2.5.1 The recommended criteria and introductory text in this chapter are intended as a guide only. The scope of this circular does not allow an exhaustive treatment of language testing. Neither this circular nor Doc 9835 is intended to replace the more extensive language testing standards, guidelines, and principles of ethics and good practice that can be found in the literature on language testing.

2.5.2 Language testing is a specialized discipline. Expert professional input is recommended at every level of aviation language test implementation and selection but is essential for test development.

3. INTRODUCTION TO LANGUAGE TESTING

3.1 Language testing standards

3.1.1 Information about generic international language testing standards can be found on the websites of a number of testing associations such as:

a) Association of Language Testers in Europe (ALTE) — http://www.alte.org; and

b) International Language Testing Association (ILTA) — http://www.iltaonline.com

3.1.2 However, it is important to recognize that existing academic or general-purpose language tests are not appropriate for the specialized domain of aviation language testing. The specific needs for aviation language testing are described below.

3.2 High stakes

3.2.1 A number of factors make language proficiency testing for compliance with Annex 1 licensing requirements a case of exceptionally high-stakes testing. Inadequate aviation language testing could result in either serious safety gaps or highly negative social and economic consequences.

3.2.2 The results of language testing seriously impact both individuals and organizations. A pilot or controller operating internationally who does not demonstrate compliance with the ICAO language proficiency requirements may be denied a licence to operate internationally, a consequence that may severely impact the career of that individual, as well as the staffing requirements of the airline or air traffic service provider for whom the individual works.

3.2.3 In addition, the safety of airline passengers depends, among other issues, on the effectiveness of pilot and air traffic controller communications. Efficient transfer of operational information is vital. When the language used in radiotelephony communications is English, then reliable, effective, and valid testing systems are required to ensure that pilots and controllers have adequate levels of English language proficiency.

3.2.4 Finally, there are economic factors to consider. State authorities, airlines, and service providers have no funds to waste on inadequate or unproven tests, neither can they afford to lose otherwise competent staff as an outcome of inadequate testing. Ultimately, they cannot afford accidents attributable to ineffective pilot/controller communication.

3.2.5 For all of these reasons, it is vital that language testing for licensing purposes comply with best practices and address the specific requirements of aviation operations.

3.3 Fundamental principles of language testing

3.3.1 There are, however, well-established principles and practices on which there is widespread professional agreement. These principles and practices, which have been incorporated into this circular, provide the recommended framework for the development and administration of aviation language tests.

3.3.2 The overriding concern of high-stakes test developers should be fairness which, in language testing, is interpreted in terms of validity and reliability. Practicality is also a fundamental test consideration. All tests should be evaluated in terms of their validity, reliability, and practicality based on documented evidence.

- *Validity.* Validity indicates the degree to which a test measures what it is supposed to measure. To this end, testers should gather and provide evidence to support the conclusions that are made about an individual's English language proficiency based on test performance.

- *Reliability.* Reliability refers to the stability of a test. Evidence should be provided that the test can be relied *upon* to produce consistent results. Reliability is usually reported in the form of a coefficient that can range from 0.0 to 1.0. Although no test will achieve a perfect reliability (1.0), tests with reliability coefficients as close to 1.0 as possible would be the most advantageous.

  There are a number of standard measures used in language test development to evaluate the reliability of a test. One method is to compare two versions of a test: the version used by one test-taker with the version used by a different test-taker. If the test is reliable, the two sets of items should be equal in difficulty and complexity. Another method of evaluating the reliability is to compare the results of a group of test-takers on a test with the results of the same group of test-takers on another established test.

  For more information about evaluating validity and reliability, refer to the document "Principles of Good *Practice* for ALTE Examinations", which is available at http://www.alte.org/quality_assurance/code/good_practice.pdf.

- *Practicality***.** Practicality refers to the balance between the resources required to develop and support a test (including the funds and the expertise) and the resources available to do so.

3.4 Test washback

3.4.1 Another important consideration related to test design is the negative or positive washback effect on training. The washback effect of testing can be described as the influence of testing on teaching and learning. It is reflected in the way trainers tend to model their curriculum around the focus areas, form and content of an examination or a test; or in the way that learners modify their learning strategies in order to succeed on a particular form of test rather than concentrating on mastering the content and skills addressed in the test.

3.4.2 A valid test designed to match the construct and content being taught (i.e. communicative language skills as defined in the rating scale) will foster a positive washback effect.

3.4.3 In contrast, an example of negative washback can be found in the older forms of the Test of English as a Foreign Language (TOEFL). The TOEFL includes a large number of discrete-point grammar questions (multiple choice or error recognition). As a result, students often neglect to learn the full range of communicative skills contained in the syllabus and prefer to spend time completing TOEFL practice tests, believing this will be an easier way to achieve a good test score. However, research indicates that such activities do not, on average, improve proficiency levels.

3.4.4 Test designers have a particular responsibility to foster positive washback as the testing process may have a considerable impact upon:

a) the validity of the test itself (are test results purely a consequence of practicing for the test, or are they a true reflection of an ability to use the language?); and

b) the way in which training is provided for the level and breadth of proficiency required to meet the standards defined by ICAO in the rating scale.

3.4.5 In summary, well-designed aviation language proficiency tests will encourage learners to focus on proficiency-building language learning activities.

3.5 Test purpose and test types

3.5.1 Tests may serve a number of different purposes which would have an influence on the test development process. Some types of common language tests and their purposes include the following:

a) *Diagnostic.* To identify strengths and weaknesses and to assess gaps.

b) *Placement.* To place students into the appropriate level of a training programme.

c) *Progress.* To measure learning progress.

d) *Achievement.* To measure what students have learned.

e) *Aptitude.* To assess an individual's ability to acquire knowledge or learn new skills.

f) *Proficiency*. To evaluate overall ability against a set of criteria.

3.5.2 A need for language testing may occur at a number of points in time in the career of a pilot or air traffic controller:

a) as a screen for pre-training selection;

b) as a diagnostic tool in a training programme;

c) as a progress check during training;

d) as a licensing requirement in fulfilment of Annex 1 requirements; or

e) as a periodic re-evaluation of proficiency.

3.5.3 The ICAO Language Proficiency Standards and Recommended Practices (SARPs) in Annex 1 require proficiency testing to fulfil licensing requirements.

3.5.4 Proficiency testing is different from progress or achievement testing in that proficiency tests do not correspond directly to a training curriculum; that is, it should not be possible for test-takers to directly prepare or study (by memorizing information, for example) for a proficiency test. Proficiency tests require test-takers to demonstrate their ability to do something representative of the full spectrum of required knowledge and skills, rather than to simply demonstrate how many of a quantifiable set of curriculum learning objectives they have learned. In an aviation context, proficiency testing should establish the ability of test-takers to effectively use appropriate language in operational conditions.

3.5.5 Proficiency tests are the only tests that are suitable for licensing purposes in the aviation community. Because licensing plays a critical role in the safety of aviation operations, this circular focuses on proficiency testing.

3.6 Delivery method

3.6.1 Speaking and listening proficiency tests can be delivered through direct or semi-direct testing. The primary difference between these two testing techniques lies in how speech samples are elicited: that is, in how the "prompts to speak" are delivered to the test-taker. Direct speaking tests involve face-to-face or telephonic interactions between the test-taker and the interlocutor, who may also serve as a rater. In semi-direct testing, test prompts and questions are pre-recorded, and test-takers' responses are recorded for evaluation at a different time and, in some cases, a different place.

3.6.2 Despite their different attributes, both live and recorded testing procedures share a common purpose: the assessment of an individual's speaking and listening abilities.

Direct testing

3.6.3 In direct testing procedures, the test-taker interacts with a "live" interlocutor, who may also be an assessor or rater. The person-to-person interaction in a direct testing procedure may be directly observed and assessed in real time by a rater or can be recorded for subsequent rating. Test-takers are asked to perform language tasks based on a set of elicitation prompts. A prompt may be a question asked by, or a topic given by, an interlocutor. The test-taker may be asked, for example, to engage in a conversation-like interview with the interlocutor or may be asked to perform in a role play.

3.6.4 One benefit of direct testing is that the test tasks can be made more natural or more communicative, as the test-takers interact with an interlocutor. Another benefit is that there is an infinite supply of test prompts available because each test is a unique interaction between the interlocutor and the test-taker. For example, if a test-taker mentions during a test that his father was an air traffic controller, the interlocutor could ask the test-taker questions related to that information — questions which the interlocutor may not ask any other test-taker. In a direct test, there is also less likelihood of a test-taker responding with rehearsed speech samples in an effort to convince an examiner of a higher level of proficiency then actually attained.

3.6.5 Direct tests require particular attention to the standardization of design and administration procedures, notably with regard to the management of time, the nature and content of language input, and overall interlocutor behaviour. This is to avoid any bias that may inadvertently arise due to the human element of the test interaction. For example, an interlocutor may, without realizing it, ask more demanding questions of one test-taker than another; or one interlocutor may speak more clearly or more slowly than another interlocutor.

3.6.6 Because direct testing requires person-to-person interactions, the administration or delivery of the test tends to be more time-consuming and human resource-intensive than semi-direct testing.

Semi-direct testing

3.6.7 In semi-direct testing, speech samples are elicited through pre-recorded and therefore standardized prompts. This is a significant benefit in that every test-taker receives the same or similar prompts, facilitating fairness. Another advantage is that the test can be administered in an audio or computer laboratory so that a larger number of test-takers can be tested at the same time.

3.6.8 However, the inflexibility arising from the use of standardized, pre-recorded prompts may limit the scope of evaluation available from semi-direct tests. This limitation may be particularly critical in the ability of the test to assess the full range of abilities covered by the "interactions" descriptors of the ICAO rating scale. Role plays and simulations conducted in this mode may be short, unnatural and restricted to the most routine aspects of language use.

Overall

3.6.9 Whether direct or semi-direct testing methods are used, it is important that test-takers are evaluated in their use of language related to routine, as well as unexpected or complicated, situations as evidence of their level of proficiency. Both direct and semi-direct tests, if well constructed, can elicit speech samples that may be assessed for proficiency in speaking and listening. Each test method has advantages and disadvantages.

4. AVIATION-SPECIFIC LANGUAGE TESTING ISSUES

4.1 Beyond the best practices of generic language testing, there are fundamental constraints specific to the ICAO language proficiency testing requirements. These concern the following:

a) test focus;

b) test content, particularly concerning the role of standardized phraseology in aviation language testing;

c) test tasks; and

d) testing for Expert Level 6 proficiency

Test focus

4.2 The ICAO language proficiency requirements focus on speaking and listening. Therefore, testing for compliance with Annex 1 licensing requirements should focus on speaking and listening proficiency.

Test content

4.3 The purpose of a language proficiency test is to assess test-takers' use of language based on their performance in an artificial situation in order to make generalizations about their ability to use language in future real-life situations. Because of the high stakes involved, pilots and air traffic controllers need to be tested in a context similar to that in which they work. Test content should, therefore, be relevant to their work roles.

4.4 Radiotelephony communication requires not only the use of ICAO standardized phraseology, but also the use of "plain" language. Phraseology is the formulaic code made up of specific words that, in the context of aviation operations, have a precise and singular operational significance. Plain language is defined in ICAO documents as "the spontaneous, creative and non-coded use of a given natural language". In simple terms, plain language can be thought of as the non-phraseology language that is used in radiotelephony communications when standardized phraseology is not appropriate.

4.5 The provisions of the ICAO language proficiency requirements that directly address test content are:

a) *Annex 1, Appendix 1,* where holistic descriptors refer to "work-related topics", "work-related context", and "routine work situation"; and

b) *Annex 1, Attachment A,* under Vocabulary and Comprehension, which refers to "work-related topics".

4.6 The use of ICAO standardized phraseology is an operational skill that is taught by qualified aviation operational specialists and is acquired to the required level of proficiency by trainee pilots and controllers during operational training. Teaching and testing standardized phraseology is an operational issue, not a language proficiency

issue. It follows that a test designed to evaluate knowledge or use of standardized phraseology cannot be used to assess plain language proficiency.

4.7 Before the ICAO language provisions were adopted in 2003, assessments of standardized phraseology were based on technical accuracy and appropriateness within the operational context and, with respect to delivery technique, only on generic "good practice". Since the adoption of the language provisions in 2003 and the publication of the language proficiency rating scale, it is recommended that assessments of ICAO standardized phraseology should, in addition to the existing guidelines in the PANS-ATM, take into account the descriptors for pronunciation and fluency of Operational Level 4.

4.8 It is acceptable that a test of plain language in a work-related context could contain a scripted test task or a prompt in which standardized phraseology is included. The test task may be used as a warm-up or as a means of setting a radiotelephony context in which to elicit plain language responses from the test-taker.

4.9 If phraseology is included in a test prompt, care should be taken that it is used appropriately and is consistent with ICAO standardized phraseology.

Test tasks

4.10 There are many kinds of test tasks or prompts that can be used to elicit speech samples. In general, tasks that resemble real-life activities are the most suitable.

4.11 It is important to keep in mind that the idea of a "work-related context" can accommodate different interpretations. A "narrow" interpretation would aim to closely replicate radiotelephony communications, including the extent of plain language needed in unusual, unexpected or emergency situations. A "broad" interpretation of the holistic descriptors and rating scale would aim to elicit plain language on various topics that are related to radiotelephony communications or aviation operations, without replicating radiotelephony communications specifically. Examples may include question and answer routines, problem-solving exchanges, briefings, simulations and role plays. Both interpretations are valid.

Testing for Expert Level 6 proficiency

4.12 The Level 6 descriptors in the ICAO rating scale refer to features of language use that go beyond the work-related context indicated in descriptors at lower levels. Formal evaluation of Level 6 using a specialized language test would follow an exhaustive procedure involving tasks and contexts that go beyond the subject matter of radiotelephony communications. Furthermore, since language proficiency at "both ends" of a proficiency scale is relatively easy to evaluate, it is not difficult to recognize "Expert" (including "native" or "native-like") proficiency. For these reasons, assessment at Level 6 should be carried out by a trained and qualified rater but not necessarily by a language testing specialist or require the use of a fully developed, specialized language test.

4.13 Monolingual native speakers of the language should be considered as "probable expert speakers". However, probable expert speakers may also include multilingual speakers who include the language as one of their "native" languages, and foreign-language speakers who have acquired a high level of proficiency. A test-taker who is tentatively considered to be a Level 6 speaker of the language may be evaluated through informal assessments (such as interviews or oral interactions with licensing authorities, recruitment officers or flight examiners), supported by documented evidence about that individual's linguistic history. This history, to be determined by state authorities, could include:

a) place of birth and early residence;

b) the language(s) used during childhood in the family, in the community and in education;

c) long periods of residence (with proven participation) in communities where the language is used socially, professionally or in education;

d) extended periods of language study or higher education diplomas; or

e) very high scores in general language tests.

4.14 On the basis of such assessment of documented information, procedures should then be documented and implemented for the formal validation of Level 6 proficiency. These procedures should be implemented and identified as assessment “events” rather than tests. They should involve a trained and qualified rater or rating team and should include assessment of language used in a work-related context with reference to the ICAO rating scale. The rater may be an operational flight or ATC examiner and the procedure may be carried out through operational assessments that include a language proficiency component.

4.15 Although the relative ease of assessing proficiency at the Expert level allows flexibility in the way such assessments may be made, the demonstration of language proficiency is nonetheless an important element of the formal process that leads toward the issuance of a pilot or an air traffic controller licence. It is therefore essential that each State establish appropriate procedures to ensure that the results of the assessment are properly documented. Because of its potential safety impact and since the outcome of a Level 6 assessment is that no further demonstration of language proficiency will be required throughout a career, the informal validation of Level 6 proficiency without documented evidence is **not** recommended.

4.16 In cases where such a procedure invalidates a suspected Level 6, the candidate may be referred to either remedial training prior to a second application of the same testing procedure and/or a formal specialized language testing procedure as described in Chapter 2. This procedure could be appropriate, for example, for native speakers whose accent or dialect is not intelligible to the aeronautical community.

———————

Chapter 2

RECOMMENDED CRITERIA FOR AVIATION LANGUAGE TESTING

The criteria listed below are formulated as self-contained statements. For personnel unfamiliar with the concepts of language testing, however, some may not be self-explanatory. Several of the criteria require documented evidence to demonstrate that they have been met. In order to facilitate the implementation of these criteria, supplementary information has been provided as described below:

— *What it means:* For some criteria, testing organizations must provide documented evidence to demonstrate that a criterion has been met. This paragraph describes the type of information required to complete an informed assessment.

— *Why it is important:* While, for language testing experts, the significance of the self-contained criterion statement may be obvious, it may not be so for personnel unfamiliar with this discipline. This paragraph justifies why a particular criterion is an essential element of testing best practices.

— *Additional information:* For several criteria, readers may require more information. This paragraph provides more explanation or links to references that may be useful.

The criteria below correlate to the item numbers in Chapter 3.

1. TEST DESIGN AND CONSTRUCT

1.1 The test should be designed to assess speaking and listening proficiency in accordance with each component of the ICAO language proficiency rating scale and the holistic descriptors in Annex 1.

— *What it means:* Language tests for flight crew and air traffic controllers should specifically address the language skills of the ICAO rating scale as well as the holistic descriptors specified in Annex 1. Testing service providers (TSPs) should be able to explain and justify their methods of and approaches to testing with evidence that all components of the ICAO rating scale have been addressed.

— *Why it is important:* The language proficiency requirements in Annex 1 specify that speaking and listening should be evaluated in the context of operational aviation communications. The holistic descriptors and rating scale were developed to address the specific requirements of radiotelephony communications, and each component of the rating scale is as important as any other. Tests developed for other purposes may not address the specific and unique requirements of aviation language testing.

— *Additional information:* The SARPs in Annex 1, Section 1.2.9, require the evaluation of the speaking and listening proficiency of pilots and air traffic controllers. Attachment A to Annex 1 provides a rating scale that describes the language proficiency levels. To test speaking and listening proficiency

requires procedures that are different from the procedures used to test reading, writing, or grammar. (See Chapter 1, 4.1.3, for some examples.) To test reading ability, knowledge about English grammar, or vocabulary items in isolation from their context is not consistent with ICAO requirements.

1.1.1 A definition of test purpose that describes both the aims of the test and the target population should be accessible to all decision-makers.

— *What it means:* Different tests have different purposes (as described in Chapter 1, 3.5) and different target populations. If an existing test is being considered, it is important that the organization offering the test clearly describe the purpose of the test and the population of test-takers for which the test is developed.

— *Why it is important:* A clear definition of test purpose and target population is a necessary starting point for evaluating the appropriateness of a test because the purpose and target population influence the process of test development and test administration. For example, a test designed to evaluate the proficiency of ab initio pilots may be very different from a test developed for experienced or professional pilots; likewise, a test designed to measure pilots' or controllers' progress during a training programme may be inappropriate as a proficiency test for licensing purposes.

1.1.2 A description of and rationale for test construct — and how it corresponds to the ICAO language proficiency requirements — should be accessible to all decision-makers in plain, layperson language.

— *What it means:* There are different approaches to proficiency testing for speaking and listening. Test developers should document the reasons for their particular approach to testing, in language that is comprehensible to people who are not experts in language test design.

— *Why it is important:* A description of the test structure and an easy-to-understand explanation of reasons for the test structure is one form of evidence that it is an appropriate tool for evaluating language proficiency according to the ICAO requirements for a given context.

— *Additional information:* See Chapter 1, 3.5 and 4.1, for more information on the issues related to aviation language testing.

1.1.3 The test should comply with principles of good practice and a code of ethics as described in Doc 9835.

— *What it means:* Doc 9835 contains a set of principles of good practice and a code of ethics, reprinted with permission from the Japan Language Testing Association (JLTA) and the International Language Testing Association (ILTA). It is important for test developers to comply with a recognized code of good practice and ethics.

— *Why it is important:* Aviation language testing is an unregulated industry and has very high stakes. A documented code of good practice and ethics, along with evidence that the organization is adhering to that code, serves as an important stopgap in an unregulated system.

— *Additional information:* In addition to the principles reprinted in Doc 9835, the Association of Language Testers in Europe (ALTE) publishes "Principles of Good Practice for ALTE Examinations", available at http://www.alte.org./quality_assurance/code/good_practice.pdf

1.1.4 The test should NOT focus on discrete-point items, on grammar explicitly, or on discrete vocabulary items. (Doc 9835)

— *What it means:* Discrete-point items are individual test questions that are presented out of context. Examples are a vocabulary test in which test-takers are asked to provide definitions for a list of words,

and a grammar test in which test-takers are asked to provide the past tense forms of a list of irregular verbs.

Discrete-point tests, also referred to as indirect tests, do not test language skills directly. Instead, they test individual, specific features of the language thought to underlie language skills; i.e., they test knowledge about grammar, vocabulary, pronunciation, etc. This type of test is not appropriate for assessing aviation language proficiency.

— *Why it is important:* The ICAO language provisions focus on the ability to use the language; discrete-point tests do not evaluate this. Furthermore, test-takers who perform well on discrete-point tests often perform poorly on tests in which they actually have to use the language.

— *Additional information:* There are a number of different ways knowledge about language is tested, for example:

a) multiple-choice questions in a series of unrelated sentences;

b) identification of an error in a sentence; or

c) written translation exercises.

For many people, such tests have the advantage of being “objective” because they give a numerical score. However, the supposed objectivity of such multiple-choice-type tests must be questioned in consideration of the choice of the particular items and questions selected for the test. It may be asked, Why were they selected from the infinite number of potential items available? In other words, Why were test-takers asked to define certain words, or why were they tested on the use of a particular tense but not on their ability to ask clarifying questions?

Speaking and listening tests, on the other hand, refer to a scale of proficiency rather than a numerical score. The rating scale describes levels of proficiency which a panel of trained raters can use to assign the test-taker a level on the rating scale.

The more directly a test performance is related to target performance, the more a test can be considered a proficiency test. For example, test administrators interested in an individual’s speaking skills should arrange for an assessment of that individual’s performance on a speaking task. Using this approach, speaking skills may be directly assessed during an interview or conversation or role play, or may be based on a recorded sample of actual speech.

The goal of a proficiency test is to assess the appropriateness and effectiveness of communication rather than grammatical accuracy. Grammatical accuracy should be considered only so far as it impacts on effective communication, but evaluating an individual’s grammatical knowledge should not be the objective of the test.

1.1.5 If comprehension is assessed through a specific listening section with individual items, it should NOT be done to the detriment of assessing interaction.

— *What it means:* Some language tests evaluate listening during an oral interaction, such as a conversation, interview or role play. Other language tests evaluate listening separately, in some cases via a series of individual listening items; one example, in the aviation language context, might require a test-taker to listen to a pre-recorded conversation between ATC and a flight crew in order to identify relevant pieces of information.

— *Why it is important:* A separate listening test can provide information about comprehension independent of a person's ability to interact. In such tests, the communication is one-way, and the test-taker does not have to "participate" in the way that is required by a conversation, role play or other interaction.

— *Additional information:* It is important for the TSP to validate the method used to evaluate comprehension.

1.1.6 Proficiency tests that are administered directly may use face-to-face communication in some phases of the delivery but should also include a component devoting time to voice-only interaction.

— *What it means:* Voice-only interaction is an important characteristic of aeronautical radiotelephony communications; when a pilot and a controller interact, they cannot see each other. Directly administered proficiency tests should simulate this condition of "voice only" in at least a portion of the test.

— *Why it is important:* When two people interact face to face, they use "non-verbal cues" (information other than words) to help them understand each other's messages. People's facial expressions, their "body language", and the gestures they make with their hands often communicate important information. Aeronautical radiotelephony communications, however, do not benefit from such non-verbal cues; all radiotelephony communication is conveyed through words alone, which can be more difficult to interpret than face-to-face communications.

— *Additional information:* In a test that is administered directly, voice-only interaction can be facilitated by means of a telephone or headset via which the interlocutor and test-taker communicate while positioned in such a way that they cannot see each other.

An appropriate strategy may be to incorporate both direct and semi-direct methods in a single testing system. In any case, the method and approach taken should be clearly justified, with evidence for the rationale of that approach provided.

1.2 The test should be specific to aviation operations.

— *What it means:* Tests should provide test-takers with opportunities to use plain language in contexts that are work-related for pilots and air traffic controllers in order to demonstrate their ability with respect to each descriptor in the rating scale and the holistic descriptors.

— *Why it is important:* The ICAO Language Proficiency Requirements (LPRs) refer to the ability to speak and understand the language used for radiotelephony communications. It is important that pilots and air traffic controllers be proficient in the plain language that they would use within the context of radiotelephony communications in order to communicate safely on any operational issue that may arise.

— *Additional information:* ICAO language provisions require proficiency in the use of standardized phraseology and in the use of plain language. The assessment of standardized phraseology is an operational activity, not a language proficiency assessment activity. While an aviation language test may include phraseology to introduce a discussion topic or make interaction meaningful to the test-taker, it is important that tests elicit a broad range of plain language and not be limited to tasks that require standardized phraseology. The focus of a language proficiency test for compliance with ICAO requirements should be on plain language.

The idea of a "work-related context" can be interpreted in different ways (Chapter 1, 4.1.10 and 4.1.11, refers). The "narrow" view would seek to replicate radiotelephony communication, including both

phraseology and plain language, as closely as possible. The “broad” view would elicit samples of interaction and comprehension on those topics occurring in radiotelephony communications without resorting to replicating radiotelephony communications. These could be of a general piloting and air traffic control nature and involve question and answer routines, short reports or problem-solving exchanges, or briefings and reports.

A further step toward providing test-takers with a familiar aviation-related context would be to customize the tests for pilots or for controllers. Thus, controllers would have the possibility of taking tests using or referring to a tower, approach or en-route environment; similarly, pilots would be able to take tests using or referring to an approach procedure. These should be seen as adaptations in the interest of the comfort of the test-taker, not as specialized tests of distinct varieties of language proficiency.

1.2.1 It is acceptable that a test contains a scripted task in which phraseology is included in a prompt, but the test should NOT be designed to assess phraseology.

— *What it means:* An aviation language proficiency test has different aims than a phraseology test. While an aviation language test can include some phraseology as prompts or “scene setters”, the purpose of the test is to assess plain language proficiency in an operational aviation context.

— *Why it is important:* First, tests of phraseology alone are not suitable for demonstrating compliance with ICAO language proficiency requirements. Second, using phraseology accurately is an operational skill which is very dependent on the operational context; and incorrect usage by a test-taker of a specific phraseology may be an operational error rather than a language error. Phraseology must be taught and tested by qualified operational personnel.

— *Additional information:* Responses containing elements of ICAO phraseology should not be rated with regard to their procedural appropriateness or technical correctness during language proficiency testing. This practice could introduce confusion between test-takers’ operational knowledge and their language proficiency. It could also introduce contradictions between the regulators’ established system of operational training/testing and language testing. Because of these contradictions, this practice could result in diminished, rather than enhanced, safety.

If phraseology is included in a test prompt, care should be taken that it is used appropriately and is consistent with ICAO standardized phraseology.

1.2.2 The test should NOT be designed to evaluate technical knowledge of operations.

— *What it means:* Language tests should not assess either operational skills or specific technical knowledge of operations. For example, a language test item may prompt the test-taker to describe an operational procedure that involves a number of steps. A test-taker may provide a very clear description of that procedure but omit one of the steps. In such a case, the rater may not recognize that the omission of one step was an operational error and may penalize the test-taker. In responding to that same test item, another test-taker may correctly identify all the steps of the process (achieving technical accuracy) but do so with problems in pronunciation and fluency based on the ICAO rating scale. In this case, the rater may, perhaps unconsciously, assign a higher level of language proficiency because of the technical accuracy than the test-taker should receive.

— *Why it is important:* If the distinction between language proficiency and technical knowledge is not clear to the interlocutor and rater of an aviation language test, it may be easy to confuse one with the other. Such confusion may lead to test-takers being penalized unfairly for technical errors or to other test-takers being rewarded, also unfairly, for their technical expertise. Another potential problem if very specific technical items are included in a language proficiency test is that they may require technical

knowledge beyond that of the test-taker; for example, answers to questions concerning ground control procedures may not be known to en-route controllers. As a result, that test-taker may be unable to respond effectively due to a lack of technical expertise rather than a lack of language proficiency.

— *Additional information:* Based on the above information, a prompt such as, “What are the separation minima for aircraft being vectored for an ILS approach?” or “Describe the different flight modes of the A320 flight control system” is, therefore, not appropriate.

1.3 The final score for each test-taker should NOT be the average or aggregate of the ratings in each of the six ICAO language proficiency skills but the LOWEST of these six ratings.

— *What it means:* For each test-taker, scores should be reported for Pronunciation, Structure, Vocabulary, Fluency, Comprehension, and Interactions in accordance with the rating scale. In cases in which a test-taker is given different ratings for different skill areas — for example, 3 for Pronunciation, 4 for Vocabulary and Structure, and 5 for Fluency, Comprehension and Interactions — the overall score for that test-taker should be the lowest of these scores; i.e., 3 in the above example.

— *Why it is important:* This practice is critical because the Operational Level 4 descriptors are developed as the safest minimum proficiency skill level necessary for aeronautical radiotelephony communications. A score lower than 4 for any one skill area indicates inadequate proficiency. For example, pilots with Operational Level 4 ratings in all areas except Pronunciation may not be understood by the air traffic controllers with whom they communicate. In summary, an individual should demonstrate proficiency to at least Level 4 in all skill areas of the ICAO rating scale in order to receive an overall Level 4 rating.

2. TEST VALIDITY AND RELIABILITY

2.1 A statement of evidence for test validity and reliability should be accessible to all decision-makers, in plain, layperson language.

— *What it means:* In language testing, fairness is interpreted in terms of validity and reliability. Validity refers to the degree a test measures what it is supposed to measure. Reliability refers to the degree that the test produces consistent and fair results. TSPs should supply documented evidence for the validity and reliability of their testing methods.

— *Why it is important:* Aviation language tests have high stakes. It is important for the safety and for the integrity of the industry, particularly for the operators and test-takers themselves, that language tests be fair and accurate. Testing systems that are not supported by documented validity and reliability may not provide, or may not seem to provide, fair and accurate results.

— *Additional information:* It is important that evidence for test validity and reliability be written in plain, layperson language. The primary target audience of documents outlining test validity and reliability should be civil aviation authority or licensing personnel rather than language testing experts. Because aviation communication safety is very much in the public interest, it is also appropriate for aviation language testing organizations to make information about the validity and reliability of their tests publicly available.

Refer to Chapter 1, Section 3.3, for more information about validity and reliability.

2.2 A description of the development process that includes the following information should be accessible to all decision-makers:

a) a summary of the development calendar; and

b) a report of each development phase.

— *What it means:* The TSP should document the entire development process.

— *Why it is important:* Before a decision is made to use a test, its quality should be examined carefully; documentation of the development process is essential to that examination. A development calendar and report will provide information about the nature and depth of analysis that went into the test development. If it is obvious that a test has been developed hastily and without the required expertise, that test may not provide, or may not seem to provide, valid and reliable results. The same is true of tests with incomplete documentation.

2.3 An appraisal of the expected test washback effect on training should be accessible to all decision-makers.

— *What it means:* Test washback refers to the effect a test has on a training programme or on students' behaviour. TSPs should demonstrate that their test will have a positive effect on training and will not encourage training that focuses on memorization and test preparation rather than on building proficiency.

— *Why it is important:* The goal of aviation operational language testing is to ensure that pilots and air traffic controllers have adequate language proficiency for the conduct of safe operations. Robust language training programmes are an essential component of a programme to enable pilots and controllers to achieve Operational Level 4 language proficiency. High-quality testing will encourage high-quality training.

— *Additional information:* Test-takers naturally will want to prepare for a test. While aviation language test-takers can memorize phraseology, they cannot acquire language proficiency as described in the LPRs simply by memorizing words and phrases. If pilots or controllers believe that certain types of narrow learning or practice activities will best and most readily prepare them for a test, they will be inclined to direct their energies to such activities, potentially at the expense of activities that can genuinely improve their language proficiency.

In the aviation environment, an example may be found in an aviation language test that focuses on the use of phraseology, to the exclusion of plain aviation language. In such a case, learners may focus their learning energies on memorizing ICAO standardized phraseology rather than on genuine language learning activities that will actually improve their English language proficiency.

Refer to Chapter 1, Section 3.4, for more information about test washback.

3. RATING

3.1 Whether rating is conducted 'live' during the assessment or after the test with recordings of the test performance, the rating process should be documented.

— *What it means:* Some speaking and listening tests rate performance during the test. Others record the test performance and rate performance later. Both rating methods are acceptable, but whichever method is used, the rating process should be explained in the test documentation.

— *Why it is important:* Because rating is one of the most important steps in language proficiency testing, it is critical to explain how it is conducted in the testing process to ensure that it is transparent to all stakeholders.

— *Additional information:* One advantage of rating test-takers after the test is that the test-taker's statements can be repeated as necessary for closer analysis by the raters. Another advantage is that the raters do not have to be physically present for the test interaction; if not physically present, raters must be able to receive an audio or video recording of the test and submit their rating reports effectively (for example, electronically). A potential advantage of rating live during the assessment may be greater efficiency.

3.2 To fulfil licensing requirements, rating should be carried out by a minimum of two raters. A third expert rater should be consulted in case of divergent scores.

— *What it means:* Best practice in language proficiency assessment calls for at least two trained and calibrated raters, at least one of whom is a language expert**.**

— *Why it is important:* Using at least two raters reduces the possibility of rater error and helps to ensure a comprehensive evaluation of each test-taker.

— *Additional information:* Ideally, an aviation language test will have two primary raters, i.e. one language expert and one operational expert — and a third rater who can resolve differences between the two primary raters' opinions. For example, there could be a situation where the primary raters agree that in five of the six skill areas a test-taker demonstrates Level 4 proficiency; however, the first rater assigns the test-taker a score of 3 on pronunciation (thereby making the test-taker's overall language proficiency level 3) and the second rater assigns the test-taker a 4 for pronunciation. A third rater would make a final determination for that skill area, and in doing so would determine the overall score for that test-taker.

A third rater would likely be involved in the process only in cases in which a test-taker obtains an overall rating of 3 or 4, since the difference between these two levels is the most critical distinction for ICAO language proficiency licensing testing.

3.3 Initial and recurrent rater training should be documented; the rater training records should be maintained, and audits of raters should be conducted and documented periodically.

— *What it means:* Language proficiency test raters need to be trained and need to be trained **together** to ensure they apply the rating scale consistently. Audits should be conducted periodically to check rater performance to ensure it is consistent over time.

— *Why it is important:* When evaluating language proficiency tests, consistency in the rating process is critical. Unlike other forms of testing in which one response to a question is correct and another response is incorrect, evaluating language proficiency relies upon subjective judgements by raters. In this context, consistency is achievable through training and experience but easy to lose without regular audits of raters and rating teams.

The reliability of test results and of the test process as a whole depends on the consistency achieved in the rating process. Audits provide a mechanism for checking consistency and, where consistency has been lost, for making adjustments as necessary.

— *Additional information:* Consistency is measured in terms of reliability. Reliability has two components:

a) *Intra-rater reliability* is the extent to which a particular rater is consistent in using a proficiency scale. In other words, does the rater apply the proficiency scale in a consistent way to all test-takers being evaluated?

b) *Inter-rater reliability* is the level of agreement between two or more independent raters in their judgement of test-takers' performance. In other words, are different raters in agreement in the scores that they assign to individual test-takers?

Raters' assessments should be monitored, both individually and comparatively, on an ongoing basis. Senior raters should formally evaluate the test raters staff periodically. Periodic cross-rating by members of different rating teams is also highly recommended as a means to prevent gradual divergence in the interpretation of the rating scale by different teams.

3.4 If rating is conducted using new technology including speech recognition technology, then the correspondence of such ratings to human rating on all aspects of the rating scale should be clearly demonstrated, in layperson language, and should be accessible to all decision-makers.

— *What it means:* If a testing organization uses new technology, such as speech recognition technology, to evaluate the speaking and listening proficiency of a test-taker, then that organization has a responsibility to clearly and plainly demonstrate that the ratings are valid and correspond to the ICAO rating scale.

— *Why it is important:* Until now, best practice in testing speaking and listening proficiency has involved the use of experienced and trained raters, who evaluate a person's proficiency based on criteria established in a rating scale. In the context of language testing, the use of speech recognition technology to evaluate human speech is a very new method, and as such, its validity and reliability should be clearly and plainly demonstrated.

— *Additional information:* The ICAO language proficiency requirements will require large-scale testing programmes. If technology can assist by making the test process easier and more cost-effective than person-by-person human rating, then it will be useful. Such testing may be particularly appropriate as a pre-test screen to determine those who may be ready for a licensing test and those who may require more training.

4. TEST ADMINISTRATION AND SECURITY

4.1 Test administration

4.1.1 A complete sample of the test should be published, including the following:

a) test-taker documents (paper instructions, screen display, etc.);

b) interlocutor instructions or prompts;

c) rater documentation (answer key, rating scale, instructions);

d) one complete sample of audio recordings (for listening sections or semi-direct prompts); and

e) demonstration of test-taker/interlocutor interaction.

— *What it means:* Decision-makers have a right to examine a complete sample of a test before they adopt, use, take, or buy it. Because of the high-stakes nature of aviation language testing, it is appropriate for testing organizations to make a complete sample of their test publicly available.

— *Why it is important:* Seeing a complete sample of a test is essential for evaluating it. Information about a test, such as a description of the test or a marketing brochure, is not sufficient for determining the test's validity, reliability, practicality, and washback effect.

— *Additional information:* It is important to note that for instructors in a training programme, being familiar with the structure and format of a test is not the same thing as "teaching to the test". Paragraph 2.3 of this chapter cautions against test designs that might provoke test-takers to try to prepare for the test by memorizing phraseology or test answers. Becoming familiar with the format of a test is good practice for both instructors and test-takers; it helps to ensure that test-takers are not unduly surprised or intimidated by the format of the test or the types of interaction it involves. For example, if the test interaction includes a voice-only segment that is conducted by telephone, it is beneficial for test-takers to be aware of this. Such knowledge does not provide them with anything they can memorize in preparation for the test; it will simply make them comfortable with the test format and the types of interaction they can expect to have during the test.

4.1.2 The test rating process should be documented, and the documentation should include instructions on the extent and nature of evidence that raters should collect.

— *What it means:* Raters should be given clear instructions on the kind of evidence they need to collect to justify and support their evaluations.

— *Why it is important:* Language is complex, and one simple statement by a person can be analysed in many different ways. Raters need to understand the depth of analysis that is expected of them in order to make and justify a rating. Documenting and supporting evaluations of test-takers are also essential in order to later review a test, either to address an appeal or complaint by a test-taker or to audit a rater or rating team (as described in 3.3 of this chapter). For such reasons, a documented set of scores alone is not sufficient; evidence and support for that score are required.

  Evidence in this context would typically include examples of language use by the test-taker that indicate strengths or weaknesses: several instances of incorrect use of verb tenses, for example, might support a particular Structure rating; or a problem pronouncing certain sounds might be documented as evidence for a Pronunciation score.

4.1.3 The instructions to the test-taker, the test administration team, and test raters should be clearly documented.

— *What it means:* Clear instructions for each part of the test process and for each stakeholder should be available and unambiguous.

— *Why it is important:* Clear instructions demonstrate that the testing organization has thoroughly considered all aspects of the testing process. Test users, test administrators, and test raters all need clear, easy-to-understand instructions if their involvement is to be effective. In addition, clear instructions help ensure that tests are administered in a consistent and therefore reliable manner.

4.1.4 The requirements for equipment, human resources, and facilities necessary for the test should be included in the instructions.

— *What it means:* The administration of tests may require a variety of equipment (computer, videotape, tape recorder), the support of different personnel (information technology personnel or sound technicians) and facilities that can accommodate the equipment and this personnel. Clear instructions for each part of the test process should be available.

— *Why it is important:* Clear descriptions of and instructions for the equipment, human resources and facilities required demonstrate that the testing organization has thoroughly considered all aspects of the testing process. Test users, test administrators, and test raters all need clear, easy-to-understand

instructions if their involvement is to be effective and to ensure that the test is administered in a consistent and therefore reliable manner.

— *Additional information:* These requirements include the room where the test will be conducted, furniture, equipment for playing any audio prompts used during the test, headsets (if used), and any other resources required by the test.

4.1.5 The testing location should be moderately comfortable, private and quiet.

— *What it means:* The testing location should not be uncomfortable or noisy.

— *Why it is important:* TSPs have an obligation to ensure a fair outcome to the test. This obligation includes eliminating undue distractions during the test.

— *Additional information:* Examples of inappropriate locations would be a staff kitchen, cafeteria, coffee lounge or hallway where people are gathering and talking. Such settings could violate the test-taker's privacy and potentially introduce distractions during the test. Similarly, a testing room that is extremely cold or hot could introduce an artificial and distracting condition to the test that could impact the test-taker's performance.

4.1.6 A full description of test administration policies and procedures should be available to all decision-makers. This description would include information about:

a) policies and procedures for retaking the test;

b) score reporting procedures (who receives the results of tests?);

c) record-keeping procedures;

d) plans for quality control, test maintenance, and ongoing test development; and

e) purchasing conditions.

— *What it means:* Policies and procedures concerning scores, records, quality control, future development, and purchasing conditions need to be clear and readily available to decision-makers and test users.

— *Why it is important:* One of the considerations in test development and test selection is whether or not there is adequate infrastructure to support and maintain the test goals.

4.1.7 A documented appeals process should be established, and information about it should be available to test-takers and decision-makers at the beginning of the testing process.

— *What it means:* All testing programmes should have an appeals process. In some cases, a re-examination may be needed. Test-takers who feel their score is not accurate may request that their test be re-rated or that they have an opportunity to retake the test.

— *Why it is important:* Even if the testing process follows best practices, errors may occur. While every appeal should not be expected to result in a complete re-scoring or re-examination, the procedures for an appeal should be clearly documented so that they can be fairly applied when appropriate.

— *Additional information:* An appeals process should address, but not be limited to, issues such as:

a) extenuating circumstances that affect the test-taker's performance. Test-takers who claim that they were "having a bad day" or "were nervous" should not be allowed an appeal, since they will need to communicate in operational situations where they will be having a bad day or feeling nervous. A test-taker, however, who has suffered a family tragedy in the days prior to the test or who was ill on the day of the test should at least be considered for an appeal;

b) steps test-takers should take to initiate an appeals process and the communication that they can expect to receive during that process;

c) the period of time (for example, 30 days or 60 days) within which the employer or licensing authority commits to resolving an appeal — either in the form of a re-review of the test, a re-examination or a rejection of the appeal.

4.2 Test security

4.2.1 A full description of security measures required to ensure the integrity of the testing process should be documented and available to all decision-makers.

— *What it means:* Test security refers to the ability of the testing organization to protect the integrity of the testing process. Testing organizations should ensure that people do not have access to specific test content or questions before the test event. In addition, TSPs should ensure that test scores are kept confidential.

— *Why it is important:* The ongoing reliability, validity, and confidentiality of a language proficiency testing system will depend heavily on the test security measures that are in place.

— *Additional information:* Testing organizations should protect test-item databases and provide secure storage of scores and test materials. They should require, establish and maintain formal commitments to confidentiality and integrity from test developers, administrators, raters, information technology (IT) personnel and any other staff who are involved in any aspect of the testing process.

Other necessary security measures to prevent cheating during test administration should include:

a) no communication between test-takers;

b) no communication between test-takers and people elsewhere during the test (for example, by use of a mobile telephone);

c) no impersonation of others; and

d) no use of false identities.

Finally, security measures should ensure the authenticity of test result data, including databases and certificates.

4.2.2 In the case of semi-direct test prompts (which are pre-scripted and pre-recorded), there should be multiple versions to meet the needs of the population to be tested with respect to its size and diversity.

— *What it means:* Tests with specific pre-recorded or pre-scripted questions or prompts require multiple versions. Decision-makers need to know that there are several versions of the test to ensure security for their particular testing needs.

— *Why it is important:* Once test items are used, there is the possibility that people may repeat or share the prompts with other test-takers; this would violate the security and validity of the test.

— *Additional information:* It is not practical to prescribe the number of versions or test prompts required for any specific test situation. The determination of what is appropriate in any situation is dependent on specific circumstances. Examples of variables that impact on the number of versions are:

a) the number of test-takers;

b) the geographic and organizational proximity of the test-takers. The closer the individuals within the test-taking population, the more likely it is that they will share their testing experience with each other. If people share test information, and that same information is used in another test, test-takers have the opportunity to prepare a response for a known test prompt. This is an example of negative test washback described in Chapter 1, 3.4; and

c) the variability inherent in the test design. A test that contains very little variability in prompts (in other words, all test-takers are asked the same questions or very similar questions) will require more frequent version changes than a test in which the interlocutor can, for a particular item, ask the test-taker a variety of questions.

It is common in large testing initiatives for a testing service to use a version of a test only once before retiring it. In other cases, a testing service develops a number of versions, then recycles them randomly. Test-takers may then generally know the sorts of questions and prompts they will encounter during a test but will be unable to predict the specific questions and prompts they will encounter during a particular testing interaction.

One security measure that testing organizations may take is to always include at least one completely new prompt or question in every version. A pattern of test-takers achieving high scores on most or all test prompts or questions but ‘failing’ the new prompt may indicate a breach in test security.

4.2.3 Test items should be held in confidence and should not be published or provided to test-takers prior to the test event.

— *What it means:* Test-takers should not have access to test questions or prompts before they take the test.

— *Why it is important:* Authorities and organizations that make test items publicly available negatively impact the integrity of the testing process. Test-takers’ prior knowledge of specific test content does not allow them to “recognize and resolve misunderstandings” or to “handle successfully and with relative ease the linguistic challenges presented by a complication or unexpected turn of events” in accordance with the ICAO language proficiency requirements. This approach will lead test-takers to memorize items and responses.

— *Additional information:* As mentioned in 4.1.1 of this chapter, one sample version of the test should be provided to decision-makers, so that they are familiar with the format of the test and the general test procedures. Specific test questions or prompts from actual tests should not be available in any way.

4.2.4 A documented policy for all aspects of test security should be accessible to all decision-makers.

— *What it means:* TSPs should clearly describe in publicly available documents how the organization establishes and maintains all required aspects of test security.

— *Why it is important:* A testing process with inadequate or unknown safeguards for test security will not be recognized as generating valid results or ensuring test-taker’s confidentiality.

— *Additional information:* All test materials, including paper documents and electronic versions, should be stored securely at all times by all stakeholders involved in test administration processes.

Periodic reviews, in the form of physical inspections, should be conducted by testing management personnel to verify that security procedures, including storage of all test materials, are being followed.

4.3 Record-keeping

4.3.1 All proficiency tests of speaking ability involving interaction between test-taker and interlocutor during the test should be recorded on audio or video media.

— *What it means:* Because aviation language testing is high stakes, it is critical that test organizations maintain either video or audio recordings of all speaking tests.

— *Why it is important:* Test recordings provide a safeguard against charges of subjective judgements and unfairness. Recordings allow a:

a) review or re-rating by different raters in case of uncertainty or an appeal; and

b) confirmation of assessments in case of appeals by test-takers or their employers.

4.3.2 Evaluation sheets and supporting documentation should be filed for a predetermined and documented period of time of sufficient duration to ensure that rating decisions can no longer be appealed.

— *What it means:* In addition to preserving the actual recording of each speaking test, for each test-taker, all score sheets and supporting documentation, including electronic data, should be filed and retained for an appropriate duration of time.

— *Why it is important:* Records are important in the case of appeals, for internal analysis related to auditing, for establishing an individual training plan and for establishing recurrent testing schedules.

— *Additional information:* At a minimum, the records should be maintained through the validity period of the licence's language proficiency requirement endorsement. Annex 1, 1.2.9.7, recommends that the maximum validity period should not surpass three years for those evaluated at Level 4, and six years for those evaluated at Level 5.

4.3.3 The record-keeping process should be documented and adequate for the scope of the testing.

— *What it means:* A testing service should document how a test-taker's performance can be captured and securely stored.

— *Why it is important:* Decision-makers need to know whether the record-keeping processes are adequate.

— *Additional information:* The outcome of the operational language assessment should comprise written comments on language performance in each skill area of the ICAO rating scale as well as the test result in terms of the demonstrated level of proficiency. In case of uncertainty, documentation should include a recommendation for assessment by a specialized language test or by another rating team.

4.3.4 The score-reporting process should be documented and scores retained for the duration of the licence.

— *What it means:* The method of scoring and the persons to whom scores are reported should be clearly documented. When a test has been rated and the results documented, the process for reporting should be clear to all decision-makers.

— *Why it is important:* This practice is important to ensure that those individuals in the organization who need to know do receive test result information and to ensure that the privacy of the test-taker and the security of the information are maintained.

4.3.5 Results of testing should be held in strict confidence and released only to the test-takers, their sponsors or employers, and the civil aviation authority, unless the test-takers provide written permission to release their results to another person or organization.

— *What it means:* The licensing authority should ensure that a policy concerning the release of test results is established. The TSP should have documented procedures on how it manages record-keeping and the confidentiality of test results.

— *Why it is important:* A confidentiality policy of test results is a key measure the licensing authority should use to manage the impact of aviation language testing on the career of pilots or controllers and the safety of passengers. A TSP should provide documented evidence on how it manages confidentiality of test results through every step of the testing process, including how it intends to transmit test results to the licensing authority.

5. ORGANIZATIONAL INFORMATION AND INFRASTRUCTURE

5.1 An aviation language testing service provider (TSP) should provide clear information about its organization and its relationships with other organizations.

5.1.1 All associations or links with other organizations should be transparent and documented.

— *What it means:* In developing and administering their aviation language test, TSPs may partner with other organizations in order to enhance their credibility with the aviation community. TSPs should provide documentation on any links to other organizations.

— *Why it is important:* In any high-stakes testing environment, relationships between a TSP and other organizations can compromise the integrity of the testing process. For example, a CAA might reject a TSP because it does not follow good testing practices; subsequently, that provider could change its name or form another organization, re-package its test and sell the same testing system (which still does not conform to good testing practices) to the CAA via deceptive marketing practices.

In order to prevent this type of deception, the provider should be required to document any other names under which it is conducting or has conducted business. The CAA should, in any case, conduct inquiries into all TSPs whose services are being considered to establish their legitimacy.

A related issue concerns claims made by TSPs of their relationships with leading industry entities. TSPs might, for example, make claims such as “Our test is endorsed by the FAA”, or “Advised by NASA”. In such cases, the provider should be required to supply documentation that explains and supports the claim, and decision-makers should contact the related organization to validate the claim.

— *Additional information:* The assessment of language proficiency for the endorsement of licences is the responsibility of Contracting States. ICAO does not accredit, certify or endorse any language TSP.

5.2 If a TSP is also a training provider, there should be a clear and documented separation between the two activities.

— *What it means:* A clear separation between testing and training activities should be documented by an organization that provides both services.

— *Why it is important:* Typically in high-stakes testing situations, testing and training should be clearly separated in order to avoid conflicts of interest. For example, an organization that provides both training and testing services could award higher scores to students within its training programme, since low scores for those students could reflect badly on the training they have received. Conversely, the organization could assign lower scores to test-takers, if the additional training those test-takers would receive would result in increased revenues for the organization's training programme.

Another potential concern would be a practice of training staff also serving as interlocutors and raters in the testing process. It is never acceptable for instructors to also be testers of their own students. There is a natural inclination for instructors to develop sympathies toward some students while perhaps regarding others less favourably. Such perceptions could interfere with the objectivity that is required of testing interlocutors and raters.

5.3 The TSP should employ sufficient numbers of qualified interlocutors and raters to administer the required tests.

— *What it means:* In addition to developing tests and new test versions, it is important that testing services have enough staff members to administer and rate the tests.

— *Why it is important:* Raters and interlocutors are usually effective only five to six hours per day. After that, tester fatigue is likely to impact their effectiveness, and their interactions and ratings can become less reliable. Testing organizations should provide evidence that they have enough trained and qualified staff to manage the volume of required tests.

5.4 Documentation on how the test is maintained, including a description of ongoing test development, should be provided.

— *What it means:* A testing organization should not only plan for the development of an initial test but also plan and budget for ongoing test development.

— *Why it is important:* An effective test that is not supported by adequate ongoing test development will not remain effective for very long. In a short period of time, test-takers will be able to predict the test items they will be presented with and to memorize responses to those items.

— *Additional information:* New test versions will constantly need to be developed. Ongoing test development should also include the creation and maintenance of a database containing all questions that have appeared on each version of a test. This practice will help to ensure that test items, or whole test versions, are not accidentally recycled as subsequent versions are developed. This practice will also enable the testing team to analyse which test items were most successful in eliciting appropriate language responses from the test-taker and those which were less successful and thus develop improved tests.

6. TESTING TEAM QUALIFICATIONS

The following lists of qualifications are provided as guidance for test development, design and administration teams as well as for organizations that aim to hire TSPs. Within a testing team, the same person may combine several areas of expertise or play several roles.

The testing team should include test designers, developers, administrators, interlocutors and raters.

6.1 Familiarity with ICAO documentation

6.1.1 All members of the testing team should be familiar with the following ICAO publications:

a) Annex 1 relevant Standards and Recommended Practices;

b) Holistic descriptors (Appendix 1 to Annex 1) and the ICAO rating scale (Attachment A to Annex 1);

c) *Manual on the Implementation of ICAO Language Proficiency Requirements* (Doc 9835); and

d) ICAO Rated Speech Samples CD.

6.2 Test design and development team

6.2.1 The test design and development team should include individuals with aviation operational, language test development, and linguistic expertise. The test design and development team should possess all three types of the following expertise:

a) Operational expertise:

1) Radiotelephony experience as a pilot, air traffic controller, or aeronautical station operator

2) Experience in aeronautical operations and procedures, and working knowledge of current practices.

b) Language test development expertise:

1) Specialization in language test development through training, education or work experience

2) Working knowledge of the principles of best practice in language test development

c) Linguistic expertise:

1) Working knowledge of the principles of theoretical and applied linguistics

2) Working knowledge of the principles of language learning

3) Experience in language teaching

— *Why it is important:* A test design and development team that includes all the above types of expertise offers the best foundation for a successful test development project.

6.3 Test administration team: administrators and interlocutors

6.3.1 Test administrators and interlocutors should have a working knowledge of the test administration guidelines published by the test organization.

6.3.2 Interlocutors should demonstrate language proficiency of at least ICAO Extended Level 5 in the language to be tested, and proficiency of Expert Level 6 if the test is designed to assess ICAO Level 6 proficiency.

6.3.3 Interlocutors should have successfully completed initial interlocutor training.

6.3.4 Interlocutors should successfully complete recurrent interlocutor training at least once each year.

6.3.5 Interlocutors should have appropriate aviation operational or language testing expertise, or both.

6.4 Rater team expertise

In Chapter 2, 3.2, it is recommended that at least two raters evaluate language tests: one with a language specialist expertise and the other with an aviation operational expertise.

a) *Operational expertise.* The involvement of operational experts such as pilots, controllers, flight instructors or examiners in the rating process will add operational integrity to the process. Operationally experienced raters can also assist by making informed judgements from an operational perspective on such aspects of language use as conciseness (exactness and brevity) in speech; and intelligibility of accents and dialects that are acceptable to the aeronautical community.

b) *Language specialist expertise.* Because language testing for licensing requirements will impact the professional careers of the test-takers as well as the reputations of operators and service providers and, ultimately, the safety of passengers and flight crews, test raters should be able not only to correctly interpret the descriptors of the rating scale but also to accurately identify strengths and weaknesses in test-takers' performance. Only qualified language specialists serving as raters can identify and describe these strengths and weaknesses.

It may be true that laypersons or "inexpert raters" (people with no academic training or qualifications in language teaching or testing) can make informal judgements about language proficiency, particularly in a pass/fail sense. However, test-takers who do not "pass" a high-stakes test will demand, and will deserve, accurate information about how their performance did not meet the target performance (in this case, Level 4 language proficiency) and the areas in which they should focus their efforts to improve performance. Likewise, a detailed justification for giving a test-taker a passing score (in this case, an overall language proficiency score of 4, 5 or 6) will need to be documented and archived.

6.4.1 Raters should demonstrate language proficiency of at least ICAO Extended Level 5 in the language to be tested. If the test is designed to assess ICAO Level 6 proficiency, raters should demonstrate language proficiency to ICAO Expert Level 6.

— *What it means:* In order to credibly and effectively evaluate test-takers' language proficiency, raters should demonstrate at least the highest level of proficiency that test-takers may achieve during assessment.

— *Why it is important:* To ensure safety, pilots and air traffic controllers expect the examiners and inspectors that assess them during operational training, and periodically thereafter, to meet stringent requirements. The assessment of language proficiency should follow the same practice, given the high stakes involved. In addition, test-takers may question the validity and reliability of the test and testing process if they have doubts concerning the credibility and qualifications of the rater.

6.4.2 Raters should be familiar with aviation English and with any vocabulary and structures that are likely to be elicited by test prompts and interactions.

— *What it means:* In order to credibly and effectively evaluate test-takers' language proficiency, raters should be familiar with the vocabulary and structures that test-takers are likely to use during the test.

— *Why it is important:* Communication between pilots and controllers is highly specialized; it includes terms that are specific to aviation (e.g. approach fix and hold position) as well as "everyday" words and structures that have singular and distinctive meanings for pilots and controllers (e.g. approach and

cleared). A rater who is unfamiliar with these terms may be confused or distracted by them during a test interaction; similarly, a rater who does not understand how pilots and controllers interact with each other may have difficulty comprehending statements made by test-takers. In cases such as these, the rater may be unable to effectively evaluate the language proficiency of test-takers in this environment.

— *Additional information:* The rater training process should include an “aviation familiarity” component so that raters can comprehend — as much as their role requires — technical aspects of the language they will hear during tests.

6.4.3 Raters should have successfully completed initial rater training.

6.4.4 Raters should successfully complete recurrent rater training at least once each year.

— *Why it is important:* Initial and recurrent training aiming to standardize rater behaviour are vital to objectivity. As a language testing standard, raters should undergo approximately 40 hours of initial rater training and 24 to 40 hours of recurrent training per year.

———————

Chapter 3

CHECKLIST

1. CHECKLIST FOR AVIATION LANGUAGE TESTING

1.1 Testing service providers (TSPs) should document adherence to the ICAO Recommended Criteria for Aviation Language Testing by completing the checklist below and submitting evidence for each item on the checklist, referencing the criterion item number. Item numbers correlate to the criteria found in Chapter 2 of this document.

1. TEST DESIGN AND CONSTRUCT

		Yes / No	*Notes*
1.1 The test is designed to assess speaking and listening proficiency in accordance with each component of the ICAO language proficiency rating scale and the holistic descriptors in Annex 1.			
1.1.1	Is a definition of the test purpose that describes both the aims of the test and the target population accessible to all decision-makers?		
1.1.2	Are a description of and rationale for test construct — and how it corresponds to the ICAO language proficiency requirements — accessible to all decision-makers in plain, layperson language?		
1.1.3	Does the test comply with principles of good practice and a code of ethics as described in Doc 9835?		
1.1.4	Does the test focus on discrete-point items, on grammar explicitly, or on discrete vocabulary items?		
1.1.5	Is a specific listening section with individual items included? *Note.— If comprehension is assessed through a specific listening section with individual items, it should NOT be done to the detriment of assessing interaction.*		
1.1.6	Does the test include voice-only interaction?		

		Yes / No	Notes
1.2 The test is specific to aviation operations.			
1.2.1	Does the test assess plain language proficiency in an aviation context?		
1.2.2	Does the test avoid items that are designed to elicit highly technical or very context-specific language?		
1.3 All six ICAO skill area criteria are assessed and reported.			
	Is the final score for each test-taker the lowest of the scores on each of the six ICAO language proficiency skills?		

2. TEST VALIDITY AND RELIABILITY

		Yes / No	Notes
2.1	Is a statement of evidence for test validity and reliability accessible to all decision-makers, in plain, layperson language?		
2.2	Is a description of the development process that includes the following information accessible to all decision-makers? • a summary of the development calendar • a report of each development phase		
2.3	Is an appraisal of the expected test washback effect on training accessible to all decision-makers?		

3. RATING

		Yes / No	Notes
3.1	Is the rating process documented?		
3.2	To fulfil licensing requirements, do at least two raters participate in the rating of tests, with a third expert rater consulted in case of divergent scores?		

		Yes / No	*Notes*
3.3	Are initial and recurrent rater training documented? Are rater training records maintained? Are raters audited periodically and reports documented?		
3.4	If rating is conducted using new technology including speech recognition technology, then is the correspondence of such ratings to human rating on all aspects of the rating scale clearly demonstrated, in layperson language?		

4. TEST ADMINISTRATION AND SECURITY

		Yes / No	*Notes*
4.1 Test administration			
4.1.1	Is a complete sample of the test published, including the following? • test-taker documents (paper instructions, screen display, etc.) • interlocutor instructions or prompts • rater documentation (answer key, rating scale, instructions) • one complete sample of audio recordings (for listening sections or semi-direct prompts) • demonstration of test-taker/interlocutor interaction		
4.1.2	Is the test rating process documented, including instructions on the extent and nature of evidence that raters should collect?		
4.1.3	Are the test instructions to the test-taker, the test administration team, and test raters clearly documented?		
4.1.4	Are the requirements for equipment, human resources, and facilities necessary for the test included in the instructions?		
4.1.5	Is the testing location moderately comfortable, private and quiet?		

		Yes / No	*Notes*
4.1.6	Is a full description of test administration policies and procedures available to all decision-makers? Does it include the following? • possibilities for retaking the test • score reporting procedures • record-keeping arrangements • plans for quality control, test maintenance, and ongoing test development • purchasing conditions		
4.1.7	Has an appeals process been established, documented and made available to test-takers and decision-makers at the beginning of the testing process?		
4.2 Test security			
4.2.1	Is a full description of security measures required to ensure the integrity of the testing process documented and available to all decision-makers?		
4.2.2	In the case of semi-direct prompts, are there multiple versions of the test to meet the needs of the population to be tested with respect to its size and diversity?		
4.2.3	Are test questions and prompts held in confidence and not published or in any way provided to test-takers prior to the test event?		
4.2.4	Is a documented policy for all aspects of test security accessible to all decision-makers?		
4.3 Record-keeping			
4.3.1	Are all proficiency tests of speaking ability recorded on audio or video media?		
4.3.2	Are evaluation sheets and supporting documentation filed and retained until a predetermined and documented period of time of sufficient duration to ensure that rating decisions can no longer be appealed?		
4.3.3	Is the record-keeping process documented and adequate for the scope of the testing?		
4.3.4	Is the score-reporting process documented and are scores retained for the duration of the licence?		

		Yes / No	*Notes*
4.3.5	Are results of testing held in strict confidence and released only to the test-takers, their sponsors or employers, and the civil aviation authority, unless the test-takers provide written permission to release their results to another person or organization?		

5. ORGANIZATIONAL INFORMATION AND INFRASTRUCTURE

		Yes / No	*Notes*
5.1	Has the aviation language TSP provided clear information on its organization and all relationships with other organizations?		
5.2	If a TSP is also a training provider, is there a clear and documented separation between the two activities?		
5.3	Does the TSP have sufficient numbers of qualified interlocutors and raters to administer the required tests?		
5.4	Has the TSP provided an explanation of how the test is maintained, including an explanation of how ongoing test development is conducted?		

6. TESTING TEAM QUALIFICATIONS

		Yes / No	*Notes*
6.1 Familiarity with ICAO documentation			
6.1.1	Are all testing team members familiar with the following ICAO publications? • Annex 1 relevant SARPS and Recommended Practices • Holistic descriptors (Appendix 1 to Annex 1) and ICAO rating scale (Attachment A to Annex 1) • *Manual on the Implementation of ICAO Language Proficiency Requirements* (Doc 9835) • ICAO Rated Speech Samples CD		

		Yes / No	*Notes*
6.2 Test design and development team			
6.2.1	Does the test design and development team include individuals with aviation operational, language test development, and linguistic expertise?		
6.3 Test administrator and interlocutor			
6.3.1	Do test administrators and interlocutors have a working knowledge of test administration guidelines?		
6.3.2	Do interlocutors demonstrate language proficiency of at least ICAO Extended Level 5 in the language to be tested, and Expert Level 6 if the test is designed to assess ICAO Level 6 proficiency?		
6.3.3	Have interlocutors successfully completed initial interlocutor training?		
6.3.4	Have interlocutors successfully completed recurrent interlocutor training at least once each year?		
6.3.5	Do interlocutors have appropriate aviation operational or language testing expertise, or both?		
6.4 Rater team expertise			
6.4.1	Do raters demonstrate language proficiency of at least ICAO Extended Level 5 in the language to be tested, and Expert Level 6 if the test is designed to assess ICAO Level 6 proficiency?		
6.4.2	Are raters familiar with aviation English and with any vocabulary and structures that will likely be elicited by the test prompts and interactions?		
6.4.3	Have raters successfully completed initial rater training?		
6.4.4	Have raters successfully completed recurrent rater training at least once each year?		

GLOSSARY OF LANGUAGE PROFICIENCY AND LANGUAGE TESTING TERMS AND ACRONYMS

Accent. A distinctive pronunciation of a language which is usually associated with a geographical region (for first language speakers) or with the phonological influence of another mother tongue (for second or foreign language speakers). All speakers of all languages have an accent.

Administration. The date or period during which a test takes place.

OR

The actions involved in the delivery of a test to a group of test-takers under specified conditions. Specifications might include registration procedures, instructions for test-taker seating arrangements, equipment needed, and time parameters for each test task.

Cue. The spoken input from an audio recording or a live interlocutor which requires the test-taker in an oral test to provide a spoken response.

Descriptor. A brief description of an aspect of language that accompanies a band on a rating scale, which summarizes the degree of proficiency or type of performance expected of a test-taker to achieve that particular score. The band may contain several descriptors.

Dialect. A distinctive variety of a language, usually associated with social or geographical distinctions, which is characterized by differences in accent, vocabulary and grammar with regard to other varieties of the same language.

Discrete item OR Discrete-point item. A test item which is not linked to any other item in the same test.

Indirect language test. A test which measures the ability or knowledge that underlies the skill that the test is intended to evaluate. An example might be testing the learners' pronunciation ability by asking them to match words that rhyme with each other or presenting written multiple choice questions as a way of testing their grammar skills.

Interlocutor. A suitably qualified and trained person with whom a test-taker interacts during a test in order to complete a speaking task.

Inter-rater reliability. The consistency or stability of scores between different raters.

Intra-rater reliability. The consistency or stability of scores given by a single rater to the same performances at different moments in time.

Item. Each testing point in a test which is given a separate mark.

Language proficiency skills. The knowledge and abilities that impact on the capacity of a given individual to communicate spontaneously, accurately, intelligibly, meaningfully and appropriately in a given language.

Note.— Six individual skills are identified in the ICAO rating scale.

Layperson. Someone who does not have special knowledge of a subject (such as language testing).

Operational language assessment. (A term specific to Doc 9835). The assessment of language proficiency using a procedure developed for a different purpose (for example, during a flight check or ATC exam). Such assessments, however, should be carried out in accordance with recognized principles of language testing best practice.

Plain language. The spontaneous, creative and non-coded use of a given natural language.

NOTES:

1) Plain language shall be used "only when standardized phraseology cannot serve an intended transmission". (Annex 10, Vol II, 5.1.1.1).

2) The choice of the term "plain" originated from existing ICAO documentation at the time of the formulation of language proficiency requirements and was preferred to other test-taker terms such as "general", "common", "extended" or "natural".

3) There is no intended association of this usage with the "Plain English" movement in the United Kingdom and the United States which aims to provide an alternative to unnecessarily complicated language by government, business, and other authorities.

Prompt. A test item or question that requires the test-taker to respond.

Rate. To assign a score or mark to a test-taker's performance in a test using a subjective assessment.

Note.— The potential for unreliability caused by individual subjectivity is countered by providing initial and recurrent training of raters, regular reference to a standard rating scale and the use of multiple raters.

Rater OR Assessor. A suitably qualified and trained person who assigns a score to a test-taker's performance in a test based on a judgement usually involving the matching of features of the performance to descriptors on a rating scale.

- ***Language rater OR Language assessor.*** A rater/assessor whose assessment will evaluate the linguistic features of a test-taker's performance in a test (compare with "operational rater").

- ***Operational rater OR Operational assessor.*** A rater/assessor whose assessment will evaluate not only the linguistic features of a test-taker's performance but also the appropriateness of a test-taker's performance in a test with regard to professional standards and procedures (compare with "language rater/language assessor").

  Note.— Knowledge of operational procedures is not tested in language tests.

Rating scale. A scale consisting of several ranked categories used for making judgements about performance. They are typically accompanied by band descriptors which make their interpretation clear.

Reliability. The consistency or stability of the measures from a test.

Response. The test-taker's linguistic performance elicited by the input of a test item (e.g. an answer to a question).

Score. The numerical or coded result of a test-taker's performance in a test enabling comparisons to be made with regard to other test-takers of the same test or with regard to a fixed standard.

Specialized language testing. (A term specific to Doc 9835). The assessment of language proficiency using a procedure that has been developed for that purpose alone and in accordance with recognized principles of language testing best practice.

Test administrator. Person who supervises and manages the administration of tests.

Test construct. A hypothesized ability or mental trait that cannot necessarily be directly observed or measured, for example, listening ability. Language tests attempt to measure the different constructs that underlie language ability.

Test maintenance. The activities of a testing organization intended to preserve the reliability, validity and security of the test over time. These activities include monitoring test results and rater reliability, designing and trialling new test items, issuing new versions of the test, and reviewing instructions for test administrators.

Test objective. The language behaviours that a test requires test-takers to demonstrate.

Test-taker OR Candidate. The person who is tested.

Test task. The combination of a single set of instructions given to candidates to guide their responses to a particular task and the associated cues and responses.

Test user. The persons or institutions making use of a test and to whom test results are made available in order for them to make informed choices, decisions or actions.

Testing system. A combination of all necessary components for administering a given test, including the test materials, and the organization of test maintenance, test delivery, rating, scoring and marking.

Validate. To undertake actions during test development and test maintenance that demonstrate the validity of a test.

Validity. The extent to which scores on a test enable inferences to be made about language proficiency which are appropriate, meaningful and useful, given the purpose of the test.

Washback effect. The influence of the format or content of tests or examinations on the methods and content of teaching and learning leading up to the assessment.

ACRONYMS

ALTE	Association of Language Testers in Europe
ILTA	International Language Testing Association
IT	Information technology
JLTA	Japan Language Testing Association
LPR	Language proficiency requirement
SARPs	Standards and Recommended Practices
TOEFL	Test of English as a Foreign Language
TSP	Testing Service Provider

— END —

ISBN 978-92-9231-271-8