د ملکي هوايي چلند اداره هـوانـوردی ملکـی
CIVIL AVIATION AUTHORITY
Islamic Republic of Afghanistan
Civil Aviation Authority

English Language Proficiency Manual (ELPT)

June 2019

Revision 0.1

Mohammad Qasim Wafayezada
Director General
Civil Aviation Authority

Approved:

د افغانستان اسلامی جمهوری دولت
ISLAMIC REPUBLIC OF AFGHANISTAN
اداره مستقل هوانوردی ملکی
AFGHANISTAN CIVIL AVIATION AUTHORITY
FLIGHT SAFETY DIRECTORATE

[This page intentionally left blank.]

CHAPTER 15
TABLE OF CONTENTS

[This page intentionally left blank.]

Chapter 15 Performance Tests: Skill and Language Proficiency

Section 15.1 Background

15.1.1 General

The examinations section of a Personnel Licencing Office (PEL Office) is responsible for administering aviation knowledge tests by preparing papers, selecting the time and place and giving clear instructions to the applicant. When the PEL Office does not have the resources to develop licencing tests, the PEL Office may choose to obtain tests or test development services from an outside provider. When a PEL Office does not have facilities for administering knowledge tests on an on-going basis, a schedule of examinations by subject shall be made available to the public through either printed or electronic means. A PEL Office may designate a Test Centre[1] should it not have accommodations for knowledge testing, or the capability to serve in a timely manner, the volume of personnel applying for aviation knowledge tests. A Test Centre is typically a non-governmental facility under the direct oversight of a PEL Office. Test centres provide applicants with a ACAA- approved testing environment for the administration of aviation knowledge examinations. PEL Office inspectors or other ACAA personnel having Test Centre surveillance responsibility may use the Checklist/Job Aid in Appendix 2 to Section 2.2 of this Chapter when conducting surveillance of the testing procedures of Test Centres. Each Test Centre and PEL Office that administers aviation knowledge tests shall have facilities for applicants that are comfortable and quiet, and provide a testing environment free from distraction.

15.1.2 ACAA PEL Office Responsibilities

The PEL Office maintains total responsibility for developing, maintaining, and administering effective performance tests. The PEL Office may choose to develop test items “in-house”. Another option is the PEL Office may choose to obtain some of these services from an outside provider (i.e., another ACAA or non-governmental service provider).

Obtaining resources or services from on outside provider does not in anyway relieve the PEL Office of its responsibility for maintaining effective tests. Development, validation, administration, and maintenance of effective tests are very complex licensing functions and if these services are obtained from an outside provider, careful oversight of the products or services obtained is of up most importance.

Section 15.2 Skill Tests

15.2.1 Criteria for Skill Test Design

The following eight criteria that should be followed in the design of a skill test:

The maximum and minimum durations of an individual test, and the proportion of time allocated to each task or element. The maximum duration has two aspects to it: the examiner cannot unnecessarily protract a test, as that may unfairly degrade the candidate's performance; and a candidate must be able to perform all practical tasks and answer all questions within a reasonable timeframe.

How minimum experience requirements are to be verified (normally, this is by reference to the candidate's logbook).

The tasks that may be re-attempted if criteria are not met, how many re-attempts may be allowed, and under what conditions. Some tasks (e.g. completing the correct actions for a practice engine failure after take-off) may be designated as "critical" and must be successfully completed on the first attempt while others (e.g. maintenance of altitude during a steep turn) may be re-attempted if criteria are exceeded on the first attempt. There should also be a limit to the overall number of re-attempts permitted.

The respective roles of examiner and candidate at all stages, particularly with respect to real or simulated emergencies. For flight tests, there must be no doubt who is pilot-in-command and the procedures for handing/taking over control of the aircraft must be clear.

The type of equipment that may be used. Most flight tests for the issue of a licence are conducted in appropriate aircraft but approved flight simulation training devices may be acceptable in some circumstances All ATC practical examinations are preferably conducted in a synthetic training device although that may not be possible or feasible for some States.

The type of assessment that is required. The most appropriate type of assessment to employ for a licence issue test is "summative evaluation", where the only intended outcome is to certify, or not, the applicant's mastery of the intended learning outcomes.

The type, content and duration of the debrief. At the conclusion of testing, the candidate should be advised of the result and, if applicable, advised of fail points or skill deficiencies, as well as aspects that were particularly well-executed. It is not normally appropriate for examiners to provide formative or diagnostic evaluation (training feedback). However, the candidate should be encouraged to self-criticise his performance and to provide feedback on the conduct of the test. examiner is not employed by the Licensing Authority. The candidate should be given a duplicate copy of the report.

The examiner report. The information that is to be recorded should be detailed as well as how the form should be processed, particularly if the examiner is not employed by the ACAA. The candidate should be given a duplicate copy of the report.

15.2.2 Inter-relationship of Knowledge and Skill Tests

The knowledge and skill tests are inter-related.

Together, they must test that the applicant has the knowledge and skill required by ICAO Annex 1 and the State's regulations for the particular licence or rating.

ICAO Annex 1 ultimately requires 100 percent successful demonstration of knowledge and skill for a licence or rating. It does not require the applicant to demonstrate only some of the knowledge (e.g., such as 70 percent of the required knowledge). Therefore, it is concluded that all areas of knowledge and skill must be successfully demonstrated.

During the written knowledge test, the applicant demonstrates his or her learning of the subject areas. The actual application of that learning is demonstrated during the skill test, although both oral questioning and actual performance of tasks are required by the licence.

While some individuals may obtain a perfect score on a written test, it is more likely that some questions will be answered incorrectly. To complete an applicant's learning in the areas missed, the ACAA Meravia requires the applicant to obtain additional training on those areas from an authorised instructor prior to taking the skill test. This will be further discussed in Module 9.

Further, upon attempting the skill test, the applicant must present his or her written test report, along with evidence of additional training, to the examiner, so that additional oral testing may be given to ensure that the applicant has mastered the subject area to the degree required by the licence or rating. This will be discussed further in Modules 10 and 11.

This is how a licensing system ensures that 100 percent demonstration of knowledge and skill has occurred.

15.2.3 Skill Test Standards

The purpose of Skill Test Standards (STS) is to describe the ACAA skill testing policy and procedures for each licence area and to provide the actual skill test, including oral questioning and the practical tasks that must be performed. The skill tests are very effective instruments for aviation safety and regulatory compliance.

Section 15.3 Language Proficiency

15.3.1 Fundamental Principles of Language Testing

While language testing is a specialized domain, there is no globally recognized single language testing authority, nor is there a single, universally accepted, best approach to language testing. As a result, there is some variability in the development and administration of language testing programmes.

However, there are well established principles and practices on which there is widespread professional agreement which are discussed in this section.

The overriding concern of high -stakes test developers should be fairness. In language testing, fairness is interpreted in terms of validity and reliability. All tests should be evaluated in terms of their validity, reliability and practically based on documented evidence.

Validity. Validity is 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.
Reliability. Reliability is the consistency or stability of the measures from a test.

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 of 1.0, one should look for tests with reliability coefficients as close to 1.0 as possible.

There are a number of standard measures used in language test development to evaluate the reliability of a test.

One example is to compare two versions of a test: the version used with one test-taker with the version used for a different test-taker. If the test is reliable, the two tests should be equal in difficulty and complexity.

Another method of evaluating the reliability of a test is to compare the results of a group of test-takers on one test with the results of the same group of test-takers on another test.

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).

15.3.2 ICAO Holistic Descriptors and Rating Scale and Proficiency Requirements

ICAO language proficiency requirements apply to speaking and listening proficiency only and do not address the ability to read or write.

In assessing a person's language proficiency, it is necessary to analyze individual categories of that person's language use (a discrete approach), as well as assess the person's overall ability to communicate in a relevant context (a holistic approach).

15.3.2.1 Holistic Descriptors

In terms of effective aviation communication, Annex 1 requires proficient speakers to be able to:

- communicate effectively in voice-only (telephone/radiotelephone) and in face-to-face situations;
- communicate on common, concrete and work-related topics with accuracy and clarity;
- use appropriate communicative strategies to exchange messages and to recognize and resolve misunderstandings (e.g. to check, confirm, or clarify information) in a general or work-related context;
- handle successfully and with relative ease the linguistic challenges presented by a complication or unexpected turn of events that occurs within the context of a routine work situation or communicative task with which they are otherwise familiar; and
- use a dialect or accent which is intelligible to the aeronautical community.

15.3.2.2 ICAO Rating Scale

It is important to note that the ICAO Rating Scale does not refer to “native” or “native-like” proficiency, a philosophical decision that “native” speech should not be privileged in a global context. All participants in aeronautical radiotelephone communications must conform to the ICAO proficiency requirements and there is no presupposition that first- language speakers necessarily conform. It is to be noted that, for international civil aviation operations, English has a clear role as an international language: it is a first language or widely used national language in about sixty countries and is an important second language in many more.

In addition to the holistic descriptors discussed above, a person needs to demonstrate a minimum standard of linguistic proficiency in each of the following specific categories:

- pronunciation;
- structure;
- vocabulary;
- fluency;
- comprehension; and
- interactions.

The ICAO Rating Scale describes the language proficiency of these specific categories according to the following six levels:

- Level 6 – Expert
- Level 5 – Extended
- Level 4 – Operational
- Level 3 – Pre-operational
- Level 2 – Elementary
- Level 1 – Pre-elementary

15.3.2.3 Determination of Language Proficiency Level.

An applicant’s language proficiency is determined by the lowest level achieved in any category.

For example, an applicant’s language categories might be individually assessed as follows:

Pronunciation – Level 3

Structure – Level 4

Vocabulary – Level 5

Fluency – Level 4

Comprehension – Level 5

Interactions – Level 4

That applicant's overall language proficiency would be assessed as Level 3, Pre-operational, despite having been assessed as Level 5 in two categories, because the person's pronunciation, stress, rhythm, and intonation are influenced by first language or regional variation, and frequently interfere with ease of understanding. In order to reach Level 4, training for that person should focus in improving pronunciation.

A summary of how these categories are used is contained in ICAO Doc 9835, Manual on the Implementation of ICAO Language Proficiency Requirements.

15.3.3 ICAO Language Proficiency Requirements for International Flight

The minimum level for international flight is Level 4, Operational.

It is well-know that some deterioration occurs in the language proficiency of individuals who do not use their second or foreign language for a long time, although people do not normally lose fully acquired first languages. Therefore a licence holder whose English language proficiency is below Level 6 and who does not regularly speak English is likely to experience some loss in proficiency over time and requires recurrent language testing.

ICAO Annex 1: 1.2.9.6 stipulates that individuals who demonstrate language proficiency below Expert Level 6 on the ICAO Rating Scale shall be formally evaluated at intervals in accordance with an individual's demonstrated proficiency level, as follows:

Level 4, Operational – should be evaluated at least once every three years;

Level 5, Extended – should be evaluated at least once every six years

Recurrent testing is not required of anyone, native or non-native speaker, who is able to demonstrate language proficiency at Level 6, Expert.

Appendix 1 to Section 15.3
ICAO Language Proficiency Rating Scale

LEVEL	*PRONUNCIATION* *Assumes a dialect and/or accent intelligible to the aeronautical community.*	*STRUCTURE* *Relevant grammatical structures and sentence patterns are determined by language functions appropriate to the task.*	*VOCABULARY*	*FLUENCY*	*COMPREHENSION*	*INTERACTIONS*
Expert 6	Pronunciation, stress, rhythm, and intonation, though possibly influenced by the first language or regional variation, almost never interfere with ease of understanding.	Both basic and complex grammatical structures and sentence patterns are consistently well controlled.	Vocabulary range and accuracy are sufficient to communicate effectively on a wide variety of familiar and unfamiliar topics. Vocabulary is idiomatic, nuanced, and sensitive to register.	Able to speak at length with a natural, effortless flow. Varies speech flow for stylistic effect, e.g. to emphasize a point. Uses appropriate discourse markers and connectors spontaneously.	Comprehension is consistently accurate in nearly all contexts and includes comprehension of linguistic and cultural subtleties.	Interacts with ease in nearly all situations. Is sensitive to verbal and non-verbal cues and responds to them appropriately.
Extended 5	Pronunciation, stress, rhythm, and intonation, though influenced by the first language or regional variation, rarely interfere with ease of understanding.	Basic grammatical structures and sentence patterns are consistently well controlled. Complex structures are attempted but with errors which sometimes interfere with meaning.	Vocabulary range and accuracy are sufficient to communicate effectively on common, concrete, and work-related topics. Paraphrases consistently and successfully. Vocabulary is sometimes idiomatic.	Able to speak at length with relative ease on familiar topics but may not vary speech flow as a stylistic device. Can make use of appropriate discourse markers or connectors.	Comprehension is accurate on common, concrete, and work-related topics and mostly accurate when the speaker is confronted with a linguistic or situational complication or an unexpected turn of events. Is able to comprehend a range of speech varieties (dialect and/or accent) or registers.	Responses are immediate, appropriate, and informative. Manages the speaker/ listener relationship effectively.
Operational 4	Pronunciation, stress, rhythm, and intonation are influenced by the first language or regional variation but only sometimes interfere with ease of understanding.	Basic grammatical structures and sentence patterns are used creatively and are usually well controlled. Errors may occur, particularly in unusual or unexpected circumstances, but rarely interfere with meaning.	Vocabulary range and accuracy are usually sufficient to communicate effectively on common, concrete, and work-related topics. Can often paraphrase successfully when lacking vocabulary in unusual or unexpected circumstances.	Produces stretches of language at an appropriate tempo. There may be occasional loss of fluency on transition from rehearsed or formulaic speech to spontaneous interaction, but this does not prevent effective communication. Can make limited use of discourse markers or connectors. Fillers are not distracting.	Comprehension is mostly accurate on common, concrete, and work-related topics when the accent or variety used is sufficiently intelligible for an international community of users. When the speaker is confronted with a linguistic or situational complication or an unexpected turn of events, comprehension may be slower or require clarification strategies.	Responses are usually immediate, appropriate, and informative. Initiates and maintains exchanges even when dealing with an unexpected turn of events. Deals adequately with apparent misunderstandings by checking, confirming, or clarifying.
Levels 1, 2 and 3 are on subsequent page.						

LEVEL	*PRONUNCIATION* *Assumes a dialect and/or accent intelligible to the aeronautical community.*	*STRUCTURE* *Relevant grammatical structures and sentence patterns are determined by language functions appropriate to the task.*	*VOCABULARY*	*FLUENCY*	*COMPREHENSION*	*INTERACTIONS*
Levels 4, 5 and 6 are on preceding page.						
Pre-operational 3	Pronunciation, stress, rhythm, and intonation are influenced by the first language or regional variation and frequently interfere with ease of understanding.	Basic grammatical structures and sentence patterns associated with predictable situations are not always well controlled. Errors frequently interfere with meaning.	Vocabulary range and accuracy are often sufficient to communicate on common, concrete, or work-related topics, but range is limited and the word choice often inappropriate. Is often unable to paraphrase successfully when lacking vocabulary.	Produces stretches of language, but phrasing and pausing are often inappropriate. Hesitations or slowness in language processing may prevent effective communication. Fillers are sometimes distracting.	Comprehension is often accurate on common, concrete, and work-related topics when the accent or variety used is sufficiently intelligible for an international community of users. May fail to understand a linguistic or situational complication or an unexpected turn of events.	Responses are sometimes immediate, appropriate, and informative. Can initiate and maintain exchanges with reasonable ease on familiar topics and in predictable situations. Generally inadequate when dealing with an unexpected turn of events.
Elementary 2	Pronunciation, stress, rhythm, and intonation are heavily influenced by the first language or regional variation and usually interfere with ease of understanding.	Shows only limited control of a few simple memorized grammatical structures and sentence patterns.	Limited vocabulary range consisting only of isolated words and memorized phrases.	Can produce very short, isolated, memorized utterances with frequent pausing and a distracting use of fillers to search for expressions and to articulate less familiar words.	Comprehension is limited to isolated, memorized phrases when they are carefully and slowly articulated.	Response time is slow and often inappropriate. Interaction is limited to simple routine exchanges.
Pre-elementary 1	Performs at a level below the Elementary level.	Performs at a level below the Elementary level.	Performs at a level below the Elementary level.	Performs at a level below the Elementary level.	Performs at a level below the Elementary level.	Performs at a level below the Elementary level.

Note: The Operational Level (Level 4) is the minimum required proficiency level for radiotelephony communication. Levels 1 through 3 describe Pre -elementary, Elementary, and Pre-operational levels of language proficiency, respectively, all of which describe a level of proficiency below the ICAO language proficiency requirement. Levels 5 and 6 describe Extended and Expert levels, at levels of proficiency more advanced than the minimum required Standard. As a whole, the scale will serve as benchmarks for training and testing, and in assisting candidates to attain the ICAO Operational Level (Level 4).

Section 15.4 Management of a Language Proficiency Test System

The seven phases of the management of a Language Proficiency Test System are:

Test design and construct
Test validity and reliability
Rating
Test Administration and
Security Recordkeeping
Organizational Information and
Infrastructure Testing-Team Qualification

15.4.1 Test design and construct.

15.4.1.1 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 ICAO Annex 1: 1.2.9 and Appendix 1.

Language tests for flight crews 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 and approaches to testing with evidence that all components of the ICAO Rating Scale are addressed.

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. 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.

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

Different tests have different purposes and different target populations. If an existing test is being considered, it is important that the organization offering the test clearly describes the purpose of the test and the population of test-takers for whom the test was developed.

A clear definition of test purpose and target population is a necessary starting point for evaluating the appropriateness of a test. The purpose and target population of a planned test 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.

15.4.1.3 A description 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.

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.

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 for the ICAO requirements for a given context.

15.4.1.4 The test should comply with principles of good practice and a code of ethics as described in Chapter 6 of ICAO Doc 9835.

It is important for test developers to comply with a recognized code of good practice and ethics. 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 stop gap in an unregulated system.

15.4.1.5 The test should not focus on discrete-point items, on grammar explicitly or on discrete vocabulary items.

Discrete-point items are individual test questions which 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. That is, they test knowledge about grammar, vocabulary, pronunciation, etc. This type of test is not
appropriate for assessing aviation language proficiency.

The ICAO language provisions focus on the ability to use the language. Discrete point tests do not evaluate a person's ability to use the language. Furthermore, test-takers who perform well on such tests often perform poorly on tests in which they actually have to use the language.

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 multiple-choice 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 testtakers 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 a 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 are 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 has an impact on effective communication, but evaluating an individual's grammatical knowledge should not be the objective of the test.

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

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. An example of an individual listening item, in the aviation language context, might require a test-taker to listen to a pre-recorded conversation between ATC and a flight crew to identify relevant pieces of information. 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 testtaker does not have to participate in the way that is required by a conversation, role-play or other interaction. It is important for the TSP to validate the method it uses to evaluate comprehension.

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

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. 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 do not benefit from such non- verbal cues; all radiotelephony communications are conveyed through words alone, which can be more difficult to interpret than face-to-face communication. 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.

15.4.1.8 The test should be specific to aviation operations.

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. The ICAO Language

Proficiency Requirements (LPRs) refer to the ability to speak and understand the language used for radiotelephony communications. It is important that flight crews and air traffic controllers be proficient in the use of plain language used within the context of radiotelephony communications in order to communicate safely on any operational issue that may arise.

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 testtaker, 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. The narrow view would seek to replicate radiotelephony communications 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 controlling 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 controllers or pilots. 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.

15.4.1.9 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.

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. 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. 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 the test-taker's operational knowledge and his/her 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 that it is consistent with ICAO standardized phraseology.

15.4.1.10 The test should not be designed to evaluate the technical knowledge of operations.

Language tests should not assess either operational skills or the specific technical knowledge of operations. A language test is not an operational or technical knowledge test. 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 that one step was an operational error and penalize the test-taker for that error. 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, because of the test-taker's technical knowledge the rater may, perhaps unconsciously, assign a higher level of language proficiency than the test-taker should receive.

If the distinction between language proficiency and technical knowledge is not very 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 getting penalized unfairly for technical errors; or to other test-takers getting 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 a test-taker; for example, answers to questions concerning ground control procedures may not be known to en-route controllers. As a result, the test-taker may be unable to respond effectively, due to a lack of technical expertise rather than a lack of language proficiency. 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" are therefore not appropriate.

15.4.1.11 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.

For each test-taker, scores should be reported for pronunciation, vocabulary, structure, 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; in the above example, the test-taker's overall score would be 3. This practice is critical because the Operational Level 4 descriptors are developed as the safest minimum proficiency skill level determined necessary for aeronautical radiotelephony communications. A lower score than 4 for any one skill area indicates inadequate proficiency. For example, a pilot with Operational Level 4 ratings in all areas except pronunciation may not be understood by the air traffic controllers with whom that pilot should 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.

15.4.2 Test validity and reliability.

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

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 of the validity and reliability of their testing methods. Aviation language tests have high stakes. It is important for safety and for the integrity of the industry, particularly the operators and for 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. 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.

15.4.2.2 A description of the development process should be accessible to all decision-makers.

The description should includes the following information:

a) a summary of the development calendar; and
b) a report on each development phase.

The TSP should document the entire development process. Before a decision is made to use a test, its quality should be examined carefully, and 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 was developed hastily and without the required expertise, that test may not provide, or seem to provide, valid and reliable results. The same is true of tests with incomplete documentation.

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

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 that their test will not encourage training that focuses on memorization and test preparation rather than on building proficiency. The goal of aviation operational language testing is to ensure that flight crews 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 ICAO Operational Level 4 language proficiency. High-quality testing will encourage high-quality training.

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 ICAO LPRs simply by memorizing words and phrases. If pilots or controllers think 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 paragraph 15.4.3 for more information about test washback.

15.4.3 Rating

15.4.3.1 The rating process to "rate" or make the applicant's score should be documented.

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 test documentation**.** Rating is one of the most important steps in language proficiency testing. It is critical to explain how rating is conducted in the testing process to ensure that it is transparent to all stakeholders.
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 of this method is that the raters do not have to be physically present for the test; in fact, raters can reside in an entirely different location, provided they can 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.

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

A rater or assessor is a suitably qualified and trained person who assigns a score to an applicant's performance in a test based on a judgment usually involving the matching of features of the performance to descriptors on a rating scale. Best practice in language proficiency assessment calls for at least two trained and calibrated raters, at least one of whom is a language expert. Using at least two raters reduces the possibility of rater error and helps to ensure a comprehensive evaluation of each test-taker.

Ideally, an aviation language test will have two primary raters — 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 may obtain 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.

15.4.3.3 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.

Language proficiency test raters need to be trained, and the raters 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. 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, making adjustments as necessary.

Consistency is measured in terms of reliability. Reliability has two components:

a) ***Intra-rater reliability*****.** 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 testtakers whom that rater is evaluating?

b) ***Inter-rater reliability*****.** 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-rater 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.

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

If a testing organization uses a 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. 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. The validity and reliability of such testing should be clearly and plainly demonstrated. 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 generally those who may be ready for a licensing test and those who require more training.

15.4.4 Test Administration and Security

15.4.4.1 Test Administration

(a) A complete sample of the test should be published.

This should include:

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); and

A demonstration of test-taker/interlocutor interaction.

Decision makers have a right to examine a complete sample of a test before they adopt, use, take or buy the test. 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. 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.

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." Avoid test designs that might provoke test-takers to try to prepare specifically for the test by memorizing phraseology or by memorizing 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.

(b) The test rating process should be documented, and the documentation should include instructions on the extent and nature of evidence that rates should collect.

Raters should be given clear instructions on the kind of evidence they need to collect to justify and support their evaluations. 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 6.3.4.3) . 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.

(c) **The instructions to the applicant, the test administration team and the test raters should be clearly documented.**

Clear instructions for each part of the test process and for each stakeholder should be available and unambiguous. 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 for their involvement to be effective. In addition, clear instructions are an important feature to ensure tests are administered in a consistent and therefore reliable manner.

(d) **The equipment, human resources and facilities necessary for the test should be included in the instructions.** 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 required equipment and personnel. Clear instructions for each part of the test process should be available.

Clear descriptions 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 for their involvement to be effective and to ensure that the test is administered in a consistent and therefore reliable manner. These requirements include the room where the test will be conducted, furniture, equipment for playing audio prompts used during the test, headsets (if used) and/or any other resources required by the test.

The testing location should offer moderate comfort, privacy and quiet. The testing location should not be uncomfortable or noisy. Aviation language testing is important. TSPs have an obligation to ensure a fair outcome to the test. This obligation includes eliminating undue distractions during the test. 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 testtaker'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.

(e) **A full description of test administration policies and procedures should be available to all decision-makers.** Policies and procedures concerning scores, records, quality control, future development, and purchasing conditions need to be clearly and readily available to decision-makers and test users. Policies and procedures concerning scores, records, quality control, future development, and purchasing conditions need to be clearly and readily available to decision-makers and test users. One of the considerations in test development and/or test selection is whether or not there is adequate infrastructure to support and maintain the test goals.

(f) **A documented appeals process should be established, and information about it should be available to applicants and decision-makers at the beginning of the testing process. All testing programmes should have an appeals process**. In some cases, a re-examination may be needed. Applicants who feel their scores are not accurate may request that their tests be re-rated or that they have the opportunity to take the test again. 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.

An appeals process should address issues such as, but not limited to:

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 when they are having a bad day or feeling nervous. But a test-taker who suffers a family tragedy in the days prior to the test, or who is ill on the day of the test, should be at least 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 reexamination
or a rejection of the appeal.

15.4.4.2 Test Security

(a) A full description of security measures required to ensure the integrity of the testing process should be documented and available to all decision-makers. 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, test center personnel should ensure that test scores are kept confidential. The ongoing reliability, validity and confidentiality of a language proficiency testing system will depend heavily on the test security measures that are in place.

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 personnel and any other staff who are involved in any aspect of the testing process. Security measures should ensure the authenticity of test result data, including databases and certificates

Other necessary security measures during test administration should prevent:

a) communication between test-takers;

b) communication between test-takers and people elsewhere during the test (for example, by use of a
mobile telephone);

c) impersonation of others; and

d) the use of false identities.

(b) **In the case of semi-direct test prompts (which are pre-scripted and pre-recorded) there should be adequate versions to meet the needs of the population to be tested with respect to its size and diversity**. Test with specific pre-recorded or pre-scripted questions or prompts require multiple versions. Decision-makers need to know that there are adequate versions of the test to ensure security for their particular testing needs.

Once test items have been 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. It is not practical to prescribe the number of versions or test prompts required for any specific test situation. The determination of what is adequate in any situation is dependent on specific circumstances. Examples of variables that impact adequacy 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; 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.

(c) **Test questions and prompts should be held in confidence, and not be published or provided to applicants prior to the test event**. Test-takers should not have access to test questions or prompts before they take the test.

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" and 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. 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.

(d) **A documented policy for all aspects of test security should be accessible to all decision-makers.** Test center personnel should clearly describe in publicly available documents how they establish and maintain all aspects of test security. A testing process with inadequate or unknown safeguards for test security will not be recognized as generating valid results or ensuring a test-taker's confidentiality. 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.

15.4.5 Record Keeping

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

Because of the high-stakes nature of aviation language testing, it is critical that test organizations maintain either video or audio recordings of all speaking tests. . Test recordings provide a safeguard against charges of subjective judgements and unfairness. Recordings allow:

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.

15.4.5.2 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.

In addition to preserving the actual recording of each speaking test, for teach applicant, all score sheets and supporting documentation, including electronic data, should be filed and retained for an appropriate duration of time. 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. At a minimum, the records should be maintained through the validity period of the licence's language proficiency endorsement requirement. Annex 1, Chapter 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.

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

A testing service should document how an applicant's performance can be captured and securely stored. Decision-makers need to know if the record-keeping processes are adequate. 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.

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

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. This practice is important to ensure that those individuals in the organisation who need to know receive test result information and to ensure that the privacy of the test-taker and the security of the information are maintained.

15.4.5.5 Results of testing should be held in strict confidence and released only to applicants, their sponsors or employers, and the ACAA, unless applicants provide written permission to release their results to another person or organization.
The ACAA should ensure that a policy concerning the release of test results is established. The test center personnel should have documented procedures on how it manages record-keeping and the confidentiality of test results.

The TSP should have documented procedures on how it manages recordkeeping and the confidentiality of test results. A confidentiality policy on test results is a key measure the licensing authority should use to manage the impact of aviation language testing on the career of a flight crew or controller 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.

15.4.6 Organisational Information and Infrastructure

15.4.6.1 An aviation language testing service provider should provide clear information about its organization and its relationships with other organizations.

15.4.6.2 All associations or links with other organisations should be transparent and documented.

In developing and administering their tests, the TSP may partner with other organisations in order to enhance their credibility with the aviation community. TSPs should provide documentation of any and all organizational links to other organizations.

In any high-stakes testing environment, relationships between a TSP and other organizations can compromise the integrity of the testing process. For example, a ACAA 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 ACAA 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 business or has conducted business in the past. The ACAA should, in any case, conduct inquiries into all TSPs whose services are being considered in order to establish their legitimacy. A related issue concerns claims made by TSPs about their relationships with leading industry entities. TSPs might, for example, make claims such as “Our test is endorsed by FAA” or “Advised by NASA.” In such cases, the provider should be required to supply documentation that explains and supports the claim, and the decision-makers should contact the related organization to validate the claim. The assessment of language proficiency for the endorsement of licences is the responsibility of the ACAA. At present, ICAO does not accredit, certify or endorse language TSPS.

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

A clear separation between testing and training activities should be documented by an organization that provides both services. Typically in high-stakes testing situations, testing and training should be clearly separated in order to avoid conflicts of interest. Two examples of conflicts of interest follow. An organization that provides both training and testing services could award higher scores to students in 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 additional training for those test -takers would result in increased revenues for the organization's training programme. Another concern regarding organizations that provide both training and testing services is the potential for training staff to also serve 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 interlocutors and raters in the testing process.

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

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. Raters and interlocutors administering or evaluating speaking proficiency tests are usually effective only five to six hours per day. After that, tester fatigue is likely to have an impact on their effectiveness, and their interactions and ratings may become less reliable. Testing organisations should provide evidence that they have enough trained and qualified staff to manage the volume of required tests.

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

A testing organization should plan not only or the development of an initial test, but is should plan and budget for ongoing test development.

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 memorize responses to those items. 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 that were less successful and thus develop improved tests subsequently.

15.4.7 Testing Team Qualifications

15.4.7.1 All Team Members.

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

- Relevant SARPS in ICAO Annex 1
- Holistic descriptors (Appendix 1 to Annex 1) and the ICAO Rating Scale (attachment A to Annex 1)
- ICAO Doc 9835; and
- ICAO Rated Speech Samples CD-ROM.

15.4.7.2 Test Design and Development Team.

The test design and development team should include individuals with specific expertise in:

- Operations;
- Language test development;

and
- Linguistics.

A test design and development team that includes all the above types of expertise offers the best foundation for a successful test development project.

15.4.7.3 Test Administration Team.

The Test administration team is made up of test administrators and interlocutors.

- Test administrators (the people who supervise and manage the administration of tests) and interlocutors should have a working knowledge of the test administration guidelines published by the test organization;
- Interlocutors should

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

  have successfully completed initial interlocutor training;

  successfully complete recurrent interlocutor training at least once a year;

  have appropriate aviation operational or language testing expertise, or both.

15.4.7.4 Raters.

(a) It is recommended that at least two raters should evaluate language tests: one with operational expertise and the other with language specialist expertise.

- **Operational expertise**. The involvement of operational experts such as pilots, controllers and 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.
- **Language specialist expertise**. Because language testing for licensing requirements will have an impact on 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 a test-taker's 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, testtakers 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, detailed justifications 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.

(b) The 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 at ICAO Expert Level 6.
- Should be familiar with aviation English and with any vocabulary and structure that are likely to be elicited by test prompts and interactions;
- Have successfully completed initial rater training;
- Successfully complete recurrent rater training at least once each year.

In order to credibly and effectively evaluate test-takers' language proficiency, raters should at least demonstrate the highest level of proficiency that test takers can achieve during assessment. 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.

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

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

(d) Raters should have successfully completed initial rater training.

(e) Raters should successfully complete recurrent rater training at least once each year.

Initial and recurrent training aiming to standardize rater behaviour is 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.

Appendix 1 to Section 15.4 Licensing Authority Audit Checklist for an ICAO Language Proficiency Test

The following checklist example is taken from Appendix C to ICAO Doc 9835 and can be used to evaluate a language proficiency test. Further guidance is available from ICAO Doc 9835.

1.0 Test design and Construct			
	Item	*Reply*	*Notes*
1.1	Is the test 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?	☐ YES ☐ NO	
1.2	Is a definition of the test purpose that describes both the aims of the test and the target population accessible to all decision-makers?	☐ YES ☐ NO	
1.3	Is 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?	☐ YES ☐ NO	
1.4	Does the test comply with principles of good practice and a code of ethics as described in Chapter 6 of ICAO Doc 9835?	☐ YES ☐ NO	
1.5	Does the test focus on discrete-point items, on grammar explicitly or on discrete vocabulary items?	☐ YES ☐ NO	
1.6	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.*	☐ YES ☐ NO	
1.7	Does the test include voice-only interaction?	☐ YES ☐ NO	

1.0 Test design and Construct			
	Item	*Reply*	*Notes*
1.8	Is the test is specific to aviation operations?	☐ YES ☐ NO	
1.9	Does the test assess plain language proficiency in an aviation context?	☐ YES ☐ NO	
1.10	Does the test avoid items that are designed to elicit highly technical or very context-specific language?	☐ YES ☐ NO	
1.11	Is the final score for each test-taker the lowest of the scores in each of the six ICAO language proficiency skills?	☐ YES ☐ NO	

2.0 Test validity and reliability			
	Item	*Reply*	*Notes*
2.1	Is a statement of evidence for test validity and reliability accessible to all decision-makers in plain, layperson language?	☐ YES ☐ NO	
2.2	Is a description of the development process that includes the following information accessible to all decision-makers: a) a summary of the development calendar? b) a report on each development phase?	☐ YES ☐ NO ☐ YES ☐ NO	
3.3	Is an appraisal of the expected test washback effect on training accessible to all decision-makers?	☐ YES ☐ NO	

3.0 Rating			
	Item	*Reply*	*Notes*
3.1	Is the rating process documented?	☐ YES ☐ NO	

3.0 Rating			
	Item	*Reply*	*Notes*
3.2	To fulfill 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	
3.3	a) Are initial and recurrent rater training documented?	☐ YES ☐ NO	
	b) Are rater training records maintained?	☐ YES ☐ NO	
	c) Are raters audited periodically and reports documented?	☐ YES ☐ NO	
3.4	If rating is conducted using new technology, including speech recognition technology, is the correspondence of such rating to human rating, on all aspects of the Rating Scale, clearly demonstrated in layperson language?	☐ YES ☐ NO	

4.0 Test administration and security			
Reference	*Item*	*Reply*	*Notes*
4.1 Test administration			
4.1.1	Is a complete sample of the test published, including the following:		
4.1.1.a	a) test-taker documents (paper instructions, screen display, etc.)?	☐ YES ☐ NO	
4.1.1.b	b) interlocutor instructions or prompts?	☐ YES ☐ NO	
4.1.1.c	c) rater documentation (answer key, rating scale, instructions)?	☐ YES ☐ NO	
4.1.1.d	d) one complete sample of audio recordings (for listening sections or semi-direct prompts)?	☐ YES ☐ NO	
4.1.1.e	e) a demonstration of test-taker/interlocutor interaction?	☐ YES ☐ NO	

4.0 Test administration and security			
Reference	*Item*	*Reply*	*Notes*
4.1.2	Is the test rating process documented, including instructions on the extent and nature of evidence that raters should collect?	☐ YES ☐ NO	
4.1.3	Are the test instructions to the test-taker, the test administration team and test raters clearly documented?	☐ YES ☐ NO	
4.1.4	Are the requirements for equipment, human resources and facilities necessary for the test included in the instructions?	☐ YES ☐ NO	
4.1.5	Is the testing location moderately comfortable, private and quiet?	☐ YES ☐ NO	
4.1.6	Is a full description of test administration policies and procedures available to all decision-makers? Does it include the following:		
4.1.6.a	a) policies and procedures for retaking the test?	☐ YES ☐ NO	
4.1.6.b	b) score reporting procedures?	☐ YES ☐ NO	
4.1.6.c	c) record-keeping arrangements?	☐ YES ☐ NO	
4.1.6.d	d) plans for quality control, test maintenance and ongoing test development?	☐ YES ☐ NO	
4.1.6.e	e) purchasing conditions?	☐ YES ☐ NO	
4.1.7	Has a documented appeals process been established and made available to test-takers and decision-makers at the beginning of the testing process?	☐ YES ☐ NO	
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?	☐ YES ☐ NO	

4.0 Test administration and security			
Reference	*Item*	*Reply*	*Notes*
4.2.2	In the case of semi-direct prompts, are there adequate versions of the test to meet the needs of the population to be tested with respect to its size and diversity?	☐ YES ☐ NO	
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?	☐ YES ☐ NO	
4.2.4	Is a documented policy for all aspects of test security accessible to all decision-makers?	☐ YES ☐ NO	

5.0 Record-keeping			
	Item	*Reply*	*Notes*
5.1	Are all proficiency tests of speaking ability involving interaction between the test-taker and interlocutor recorded on audio or video media?	☐ YES ☐ NO	
5.2	Are evaluation sheets and supporting documentation filed for a predetermined and documented period of time of sufficient duration to ensure that rating decisions can no longer be appealed?	☐ YES ☐ NO	
5.3	Is the record-keeping process adequate for the scope of the testing and documented?	☐ YES ☐ NO	
5.4	Is the score-reporting process documented, and are scores retained for the duration of the licence?	☐ YES ☐ NO	

5.0 Record-keeping			
	Item	*Reply*	*Notes*
5.5	Are results of testing held in strict confidence and released only to test-takers, their sponsors or employers, and the civil aviation authority, unless test-takers provide written permission to release their results to another person or organization?	☐ YES ☐ NO	

6.0 Organizational information and infrastructure			
	Item	*Reply*	*Notes*
6.1	Has an aviation language TSP provided clear information about its organization and its relationships with other organizations?	☐ YES ☐ NO	
6.2	If a TSP is also a training provider, is there a clear and documented separation between the two activities?	☐ YES ☐ NO	
6.3	Does the TSP employ sufficient numbers of qualified interlocutors and raters to administer the required tests?	☐ YES ☐ NO	
6.4	Has the TSP provided an explanation of how the test is maintained, including an explanation of how ongoing test development is conducted?	☐ YES ☐ NO	

7.0 Testing-team qualifications			
	Item	*Reply*	*Notes*
7.1 Familiarity with ICAO documentation			
7.1.1	Are all testing team members familiar with the following ICAO publications?		
7.1.1.a	a) the relevant SARPS and Recommended Practices of Annex 1?	☐ YES ☐ NO	

7.0 Testing-team qualifications			
	Item	*Reply*	*Notes*
7.1.1.b	b) holistic descriptors (Appendix 1 to Annex 1) and the ICAO Rating Scale (Attachment A to Annex 1)?	☐ YES ☐ NO	
7.1.1.c	c) *Manual on the Implementation of ICAO Language Proficiency Requirements* (Doc 9835)?	☐ YES ☐ NO	
7.1.1.d	d) ICAO Rated Speech Samples CD?	☐ YES ☐ NO	
7.2 Test design and development team			
7.2.1	Does the test design and development team include individuals with aviation operational, language test development, and linguistic expertise?	☐ YES ☐ NO	
7.3 Test administration team (administrators and interlocutors)			
7.3.1	Do test administrators and interlocutors have a working knowledge of the test administration guidelines published by the test organization?	☐ YES ☐ NO	
7.3.2	Do interlocutors demonstrate language proficiency of at least ICAO Extended Level 5 in the language to be tested and proficiency at Expert Level 6 if the test is designed to assess ICAO Level 6 proficiency?	☐ YES ☐ NO	
7.3.3	Have interlocutors successfully completed initial interlocutor training?	☐ YES ☐ NO	
7.3.4	Have interlocutors successfully completed recurrent interlocutor training at least once each year?	☐ YES ☐ NO	
7.3.5	Do interlocutors have appropriate aviation operational or language testing expertise, or both?	☐ YES ☐ NO	
7.4 Rater team			
7.4.1	Do raters demonstrate language proficiency of at least ICAO Extended Level 5 in the language to be tested,	☐ YES ☐ NO	

7.0 Testing-team qualifications			
	Item	*Reply*	*Notes*
	and Expert Level 6 if the test is designed to assess ICAO Level 6 proficiency?		
7.4.2	Are raters familiar with aviation English and with any vocabulary and structures that will likely be elicited by the test prompts and interactions?	☐ YES ☐ NO	
7.4.3	Have raters successfully completed initial rater training?	☐ YES ☐ NO	
7.4.4	Have raters successfully completed recurrent rater training at least once each year?	☐ YES ☐ NO	

[Page intentionally left blank.]